ChinaXiv:202105.00070v16

Copula 熵：理论和应用

马健

日立中国研究院
majian@hitachi.cn

摘要

统计独立性是统计学和机器学习领域的基础性概念，如何表示和度量统计独立性是该领域的基本问题。Copula 理论提供了统计相关关系表示的理论工具，而 Copula 熵理论则给出了度量统计独立性的概念工具。本文综述了 Copula 熵的理论和应用，概述了其基本概念定义、定理和性质，以及估计方法。介绍了 Copula 熵研究的最新进展，包括其在统计学的九个基本问题（结构学习、关联发现、变量选择、因果发现、系统辨识、时延估计、域自适应、正态性检验和双样本检验等）上的理论应用。讨论了前四个理论应用之间的关系，以及其对应的深层次的相关性和因果性概念之间的联系，并将 Copula 熵的（条件）独立性度量框架与基于核函数和距离相关的同类框架进行了理论对比，又通过仿真和实际数据实验评估验证了 Copula 熵的实际优越性。简述了 Copula 熵在理论物理学、理论化学、化学信息学、材料学、水文学、气候学、气象学、环境学、生态学、动物形态学、农学、认知神经学、运动神经学、计算神经学、心理学、系统生物学、生物信息学、临床诊断学、老年医学、精神病学、公共卫生学、经济学、管理学、社会学、教育学、计算语言学、新闻传播学、法学、政治学、军事学、情报学，以及能源工程、食品工程、土木建筑、交通运输、制造工程、可靠性工程、化学工程、航空航天、车辆工程、电子工程、通信工程、高性能计算、测绘遥感和金融工程等领域的实际应用。

关键词: Copula 熵；传递熵；统计独立性；条件独立性；相关性；因果性；结构学习；关联发现；变量选择；因果发现；系统辨识；时延估计；域自适应；正态性检验；双样本检验；交叉学科应用

Abstract

Statistical independence is a core concept in statistics and machine learning. Representing and measuring independence are of fundamental importance in related fields. Copula theory provides the tool for representing statistical independence, while Copula Entropy (CE) presents the tool for measuring statistical independence. This paper first introduces the theory of CE, including its definition, theorem, properties, and estimation method. The theoretical applications of CE to structure learning, association discovery, variable selection, causal discovery, system identification, time lag estimation, domain adaptation, multivariate normality test, and two-sample test are reviewed. The relationships between the former four applications and their connection to correlation and causality are discussed. The frameworks based on CE, the kernel method, and distance correlation for measuring statistical independence and conditional independence are compared. The advantage of CE over other independence and conditional independence measures is evaluated. The applications of CE in theoretical physics, theoretical chemistry, cheminformatics, materials science, hydrology, climatology, meteorology, environmental science, ecology, animal morphology, agronomy, cognitive neuroscience, motor neuroscience, computational neuroscience, psychology, system biology, bioinformatics, clinical diagnostics, geriatrics, psychiatry, public health, economics, management, sociology, pedagogy, computational linguistics, mass media, law, political science, military science, informatics, energy, food engineering, architecture, civil engineering, transportation, manufacturing, reliability, chemical engineering, aeronautics and astronautics, automobile, electronics, communication, high performance computing, remote sensing, and finance are briefly introduced.

Keywords: copula entropy; transfer entropy; correlation; causality; structure learning; association discovery; variable selection; causal discovery; system identification; time lag estimation; domain adaptation; normality test; two-sample test; multidisciplinary application

ChinaXiv:202105.00070v16

目录

ChinaXiv:202105.00070v16

ChinaXiv:202105.00070v16

1 引言

统计独立性是统计学和机器学习领域的基础性概念，如何表示和度量统计独立性是统计学的基本问题。在统计学早期的 19 世纪，就有 Pearson [1] 提出了相关系数的概念来度量统计独立性，并应用于优生学的研究。上个世纪，在对相关性的研究中 Copula 函数理论被提出，提供一种统一表示随机变量之间统计关联关系的理论工具 [2,3]。根据 Sklar 定理 [4]，通俗地讲，任何一个多变量之间的关联关系都对应着一个用于表示这种关系的函数，称为 Copula 函数。Copula 函数表示了多变量之间全部的关联关系，且与单个变量的性质是无关的。

2008 年，马健和孙增圻提出了 Copula 熵（Copula Entropy: CE）的概念 [5]。CE 的概念由 Copula 密度函数定义而来，本质上是一种香农熵的形式。我们也证明了它与信息论 [6] 中的互信息概念是等价的。事实上，CE 的提出是受到了这样的启发，Copula 函数被认为包含了全部的关联关系，而互信息一直被认为度量了全部的关联关系的信息，那么我们认为这二者之间必然有某种联系。对这种必然联系的研究的结果，就是提出了 CE 的理论。

CE 是一种多变量之间关联关系度量的理论，与关联关系表示理论——Copula 函数理论相对应。Copula 函数表示关联关系，而由之得到的 CE 度量了关系中的信息量。CE 是一个理想的统计独立性度量的概念，具有很多优美的属性，包括对称性、非正性、单调变换不变性、以及在高斯变量时与相关系数等价等。

CE 是一种理想的统计相关性度量工具，同时它又可以用来表示和度量另一个重要的统计学概念——条件独立性（Conditional Independence：CI）。这样，我们就得到了一个基于 CE 的（条件）独立性度量理论框架，将相关性和因果性这两个基本概念统一起来。

CE 是一个基础性的统计工具，可以用来解决多个统计学的基本问题。我们在 2008 年就将其应用到结构学习问题上 [7]，用来学习统计变量之间的关联关系结构。最近，我们又将其应用到关联发现 [8]、变量选择 [9]、因果发现 [10]、域自适应 [11]、正态性检验 [12]、双样本检验 [13]、时延估计 [14] 和系统辨识 [15] 等问题上，都取得了良好的应用效果。

作为一种基础性的数据分析工具，CE 被提出以来，在多个不同学科得到了实际的应用，包括理论物理学 [16]、理论化学 [17]、化学信息学 [18]、材料学 [19]、水文学 [20–43]、气候学 [44,45]、气象学 [10,46–48]、环境学 [49,50]、生态学 [51]、动物形态学 [52,53]、农学 [54]、认知神经学 [55–61]、运动神经学 [62–66]、计算神经学 [67,68]、心理学 [69]、系统生物学 [70,71]、生物信息学 [72–74]、临床诊断学 [9,75–78]、老年医学 [79–82]、精神病学 [83,84]、公共卫生学 [53,75]、经济学 [85–89]、管理学 [90,91]、社会学 [11]、教育学 [92]、计算语言学 [93]、新闻传播学 [94]、法学 [95]、政治学 [96]、军事学 [97]、情报学 [98]，以及能源工程 [14,99–107]、食品工程 [108,109]、土木建筑 [110,111]、交通运输 [112,113]、制造工程 [114–117]、可靠性工程 [118]、化学工程 [119–121]、航空航天 [122–125]、

ChinaXiv:202105.00070v16

车辆工程 [126]、电子工程 [127]、通信工程 [128,129]、高性能计算 [130]、测绘遥感 [131] 和金融工程 [132–140] 等。在这些应用中，CE 被用来分析和度量多学科数据中的统计关联性或因果性，用以增加对数据中变量间统计关系的理解，或者用于建立和评价模型。CE 工具不仅为各种应用提供了理论支撑，同时也改进了计算的可靠性和效率。

在以上实际应用中，研究者也提出了一些基于 CE 理论的新方法，如 GCMI 方法 [56]、信息瓶颈（Information Bottleneck）计算 [18]、独特信息（Unique Information）估计 [68]、非对称互信息估计 [53]、图结构相似度计算 [52]、聚类算法 [41,45]、非线性主元分析 [120]、决策树构建 [85,86]、因果压缩 [72]、因果结构学习 [87,108,109]、Copula 参数估计 [40]、藤 Copula 结构选择 [135,137] 和 Survival CE [78] 等。CE 作为一种基础性的理论方法，给出了一种处理相关性和因果性的普适性基本工具，为更多新方法论的派生提供了可能。

本文第2部分介绍 Copula 熵的理论和估计方法，第3部分介绍 CE 在统计学中的理论应用，用于解决统计学的八个基本问题，第4部分讨论四个相关的问题，第5部分简要叙述 CE 在多个不同学科领域中的实际应用，第6部分对论文进行总结。

ChinaXiv:202105.00070v16

2 Copula 熵

2.1 理论

Copula 理论是关于多随机变量之间相互依赖关系表示的理论 [2,3]。此理论定义一类函数，成为 Copula 函数，定义如下：

定义 1 (Copula 函数) *给定 N 维随机向量 $\mathbf{X} = (X_1, \ldots, X_N) \in R^N$. 令 $\mathbf{u}$ 表示 $\mathbf{X}$ 的边缘分布函数 $u_i = F_i(x_i), i = 1, \ldots, N$. 则 $\mathbf{X}$ 对应 N 维 Copula 函数 $C : I^N \to I, I = [0,1]$ 需要满足如下性质：*

1. *C 的下确界为 0 且在单位立方体内的任意子立方体内单调递增；*
2. *$C(1, \ldots, 1, u_i, 1, \ldots, 1) = u_i$.*

直观的理解，Copula 函数就是在单位 N 立方体上的分布函数，边缘分布为均匀分布，下确界为 0，且在任意向上方向上单调增加。从 Copula 函数出发，对各变量求导，可以很容易地定义与之相对应的 *Copula* 密度函数 $c(\mathbf{u})$。

Copula 理论的核心结论是 Sklar 定理，给出了如何利用 Copula 函数表示随机变量依赖关系的结论，如下：

定理 1 (Sklar 定理) *[4] 给定任意 N 维随机变量 $\mathbf{X}$ 的联合分布函数 $F(\mathbf{X})$、边缘分布函数 $F_i(X_i)$ 和 Copula 函数 $C(\mathbf{u})$，则联合分布函数可以表示为输入为边缘分布函数的 Copula 函数的形式，如下：*

$$F(\mathbf{x}) = C(F_1(x_1), \ldots, F_N(x_N)). \tag{1}$$

Copula 函数的表示将多变量的联合分布与单个变量的联合分布分离开来，将依赖关系表示为一个 Copula 函数。因此，依赖关系与单个变量的属性是没有关系的，Copula 函数中包含了全部的依赖关系信息。对式 (1)两边求导，就得到相应的 Sklar 定理的密度函数版本：

$$p(\mathbf{x}) = c(\mathbf{u})\prod_i p(x_i). \tag{2}$$

其中，$p(\cdot)$ 表示概率密度函数。

利用 Copula 密度函数的表示，我们就可以定义 Copula 熵，如下：

定义 2 (Copula 熵) *[5] 给定多随机变量 $\mathbf{X}$，及其边缘分布 $\mathbf{u}$ 和 Copula 密度函数 $c(\mathbf{u})$，则 Copula 熵定义为：*

$$H_c(\mathbf{x}) = -\int_{\mathbf{u}} c(\mathbf{u})\log c(\mathbf{u})d\mathbf{u}. \tag{3}$$

在信息论中，互信息（Mutual Information：MI）和熵是两个定义不同的概念 [6]。在文献 [5] 中，我们证明了二者本质上是相同的，也即是，MI 等价于负的 CE，也可以表示成熵的形式。定理如下：

定理 2 *多随机变量的 MI 等价于其负的 CE.*

$$I(\mathbf{x}) = -H_c(\mathbf{x}). \tag{4}$$

定理的证明很简单。由定理可以立即得到一条关于联合熵、边缘熵和 CE 之间关系的推论，如下：

推论 1 *多随机变量的联合熵等于边缘熵和 CE 的和。*

$$H(\mathbf{x}) = \sum_i H(x_i) + H_c(\mathbf{x}). \tag{5}$$

以上结论通过 CE 的定义，加深了我们对信息论基本概念及其之间关系的了解，也因此在 Copula 理论和信息论之间架起了一座桥梁。

2.2 性质

熵性 由 Copula 理论得到的 CE 具有很多有趣的性质。首先从定义来看，CE 是一种特殊的香农熵，定义在单位体的概率分布函数上，因此其也具有香农熵具有的连续性、对称性和可加性等特性。

多变量 香农的 MI 定义针对的是二变量情况，但 CE 概念不限于二变量的情况，也适用于多变量的情况，且多变量之间具有对称性，扩展了 MI 的定义和适用范围。

ChinaXiv:202105.00070v16

全阶次 由 Copula 密度函数而定义的 CE 从一个新的角度给出了对 MI 概念更深入的理解。Copula 函数被认为是包含了随机变量之间所有相关性的信息，那么 CE 作为相关性的随机性的度量，就等于给出了随机变量之间所有阶次相关性的信息量。

单调变换不变性 由于 Copula 函数具有单调变换不变性，因此基于 Copula 函数定义的 CE 天然地继承了这一不变性特性。

边缘函数无关 上面提到，Copula 理论将联合分布分解为边缘函数和 Copula 函数两个相对独立的部分，这也对应到联合熵的分解：随机变量的联合熵也可以相应地分解为边缘熵和 CE 两个相互无关的部分。而 MI 与 CE 等价，因此 MI（CE）只与 Copula 函数有关，与边缘函数无关、联合函数无关，这与香农基于边缘函数和联合函数的 MI 定义构成了显著的理论区别。

非正性 需要指出的是，CE 本身是非正的，它表明了由于多变量之间具有相关性，使得多变量之间相互包含有其他变量的信息，因此就使得联合熵的总信息量减少，表现为联合熵小于各个变量的边缘熵之和。一般地讲，熵度量了随机变量的不确定性，是非负的；而 CE 则是非正的，因为它度量了由于变量间相关性导致减少的不确定性。

等价关系 相关系数是统计学传统的相关性度量，它隐含着分布高斯性的假设。可以很容易证明，在高斯分布的情况下，相关系数与 CE 具有数学上的等价关系，即 CE 可以由相关系数矩阵来表示。

2.3 估计方法

MI 作为信息论的基本概念，具有广泛的应用价值。但学界普遍认为 MI 的估计是十分困难的。我们根据定理 2，给出了一个简单且优雅的非参数 CE（MI）估计方法[1] [5]。该方法仅需如下 2 步：

1. 估计经验 Copula 密度函数；
2. 由经验 Copula 密度函数估计 CE。

给定随机变量 $\mathbf{X}$ 的一组独立同分布样本 $\{x_1,\ldots,x_T\}$，可以很容易地通过次序统计量（rank）来估计经验 Copula 密度函数，如下

$$F_i(x_i) = \frac{1}{T}\sum_{t=1}^{T}\mathbf{1}(x_t^i < x_i), \tag{6}$$

[1]本方法已经实现为 R 和 Python 语言的 **copent** 算法包 [141]，并已分别在 CRAN 和 PyPI 上发布共享。

ChinaXiv:202105.00070v16

其中 $\mathbf{1}(\cdot)$ 表示示性函数。

在得到经验 Copula 密度函数后，第 2 步就是一个熵估计的问题，有很多方法可以采用。我们采用了 Kraskov 等 [142] 提出的 k 近邻法来估计 CE，因为它是一个非参数方法，具有良好的估计性能。

由于在两步中都采用了非参数的方法（次序统计量和 k 近邻法），因此，我们就得到了一个非参数的 CE 估计方法。方法简单，易于实现，且计算量要求较低。此方法是一个典型的基于序数（rank）统计量的非参数估计方法，将 CE 非参数估计的本质等价于计算归一化的序数统计量的熵，内涵深刻。

3 理论应用

3.1 结构学习

从数据分析一组随机变量之间的关联结构，可以帮助我们了解系统内部的内在结构关联性，具有重要的应用价值。在统计和机器学习学习中，表示这种关联结构的主要工具方法是图（Graph），图中的顶点表示随机变量，顶点之间的边表示变量之间的关联，边的权重表示关联的强度。图又分为有向图和无向图，前者的边具有方向而后者则无方向，前者表示变量之间的因果关系而后者表示关联关系。从数据中学习这种关联图结构的问题，被称为结构学习（Structure Learning）。

结构学习的算法很多，其中比较著名的有 Chow-Liu 的图结构学习方法 [143]。该方法通过学习变量的互信息矩阵，再基于矩阵学习最小生成树（Minimal-Spanning-Tree：MST）来得到主要关联结构的骨架。

利用互信息和 CE 的等价性，我们给出了 Chow-Liu 算法的 CE 版本 [7]，包含两步：

1. 利用 CE 估计方法学习得到随机变量的关联矩阵；
2. 再利用 MST 生成算法从上述矩阵得到关联图结构。

由于我们的 CE 估计方法简单有效，相较于传统的互信息估计具有明显优势，因此也使得 Chow-Liu 算法更可靠有效。

我们将算法应用到两个经典的 UCI 机器学习数据集 [144]：鲍鱼生长数据集和波士顿房价数据集。实验结果显示，算法能够得到具有可解释性的关联结构，使我们对数据反映的鲍鱼生长特性和波士顿房价相关因素的内在关系有了更深入的理解。[2]

[2]实验代码：`https://github.com/majianthu/dse`

ChinaXiv:202105.00070v16

3.2 关联发现

经验科学是分析数据的学问。通过分析收集的观察或经验数据，人们得出对象系统的科学结论。关联的概念是多元统计分析的基本工具之一。它度量了随机变量之间的统计性内在联系，进而被赋予科学意义。发现关联关系是科学研究的主要内容方法之一。

Pearson 相关系数 [1] 是一种统计学史上重要的相关性度量概念，教科书里都会讲到，应用也很广泛。但由于它是统计学早期提出的概念，因此具有很多局限性。从理论上来讲，它只适用于线性相关关系的情况，隐含着高斯分布的假设，使它在绝大多数实际情况中都不适用。它是一个二变量的度量，没有多变量的版本。

CE 则是一种更高级的相关性度量，相对于 Pearson 相关系数具有显著的优势。它没有线性和高斯性的假设，且是一个多变量的相关性度量。实际上，CE 度量的是统计独立性，比相关性更宽泛的概念，在统计独立的情况下，其为 0。CE 还具有单调变换不变性，且在高斯分布的情况下，与相关系数等价。简单列一下 CE 作为相关性度量的优点：

- 无模型假设，
- 可处理非线性关系，
- 统计独立性度量，
- 单调变换不变性，
- 在高斯情况下与相关系数等价。

综合了如此多优点，CE 是一个完美的相关性度量，完全可以替代 Pearson 相关系数，适用于任何类型的相关性度量。Pearson 相关系数作为一个历史悠久的统计工具，可以进入历史了。

关于 CE 与 Pearson 相关系数的理论上的对比，可参见论文 [8]。论文还利用著名的 NHANES 医学体检数据 [145]，从实验上证明了 CE 的显著优越性[3]。

3.3 变量选择

变量选择（Variable Selection），又称特征选择，是统计和机器学习的基本问题 [146,147]。当人们试图从一组自变量和目标预测变量之间建立函数关系时，往往希望只选择真正与目标变量有内在联系的自变量子集作为函数模型的输入，以提高模型的科学性（或可解释性），同时降低模型的复杂度。这样的问题称为变量选择。在统计和机器学习中，变量选择主要用于多元分类或回归分析中建立的函数模型关系。

[3]实验代码：`https://github.com/majianthu/nhanes`

ChinaXiv:202105.00070v16

传统的变量选择方法很多，主要的有准则法、模型正则化方法和关联度量方法。主要的准则法有 AIC [148] 和 BIC [149] 等，通过在似然函数上加上对模型复杂度的惩罚项得到。模型正则化方法主要用于广义线性回归模型，在学习模型的过程中，通过在似然函数上加上模型参数（线性系数）的 1 范数或 2 范数或二者的组合得到，经典的方法包括 LASSO [150]、岭回归（Ridge Regression）[151] 和弹性网络（Elastic Net）[152] 等。以上两类方法都是基于似然函数加惩罚项的形式完成变量选择，都是模型有关的。关联度量的方法则是通过自变量和目标变量之间的关联强度来选择变量，通常是模型无关的。主要的关联关系度量包括传统的 Pearson 相关系数，但它只能度量线性关系，仅适用于线性模型。其它几个主要的非线性关联度量也都被应用到变量选择问题上，包括希尔伯特-施密特独立性准则（Hilbert-Schmidt Independence Criterion：HSIC）[153,154] 和距离相关（Distance Correlation：DC）[155,156] 等。

变量选择问题，推荐 CE 方法，不建议 LASSO 或者 p-value 等传统统计方法。本方法利用 CE 度量自变量和目标变量之间的关联强度，根据强度从大到小依次选择变量。在变量选择问题上，CE 已被真实数据实验证明优于以下主流变量选择方法：

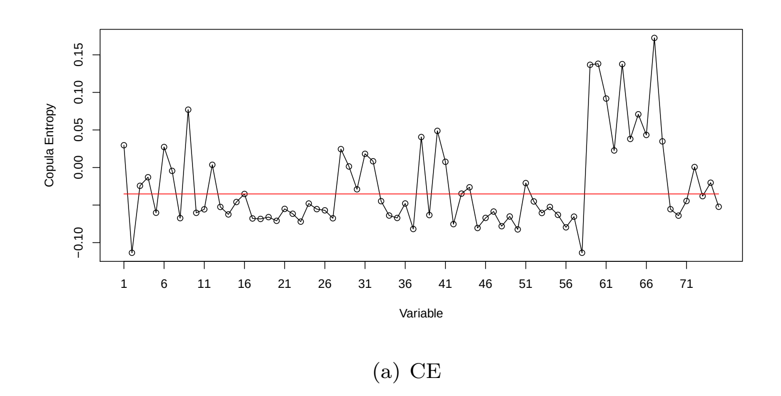
- LASSO / Ridge Regression / Elastic Net [150–152],
- AIC / BIC [148, 149],
- Adaptive LASSO [157],
- Hilbert-Schimdt Independence Criterion (HSIC) [153, 154],
- Distance Correlation [155, 156],
- Heller-Heller-Gorfine Tests of Independence [158],
- Hoeffding's D test [159],
- Bergsma-Dassios T* sign covariance [160],
- Ball correlation [161].

实验[4]采用了著名的 UCI 心脏病数据集 [144]，将 CE 方法与以上方法进行对比。该数据集包含了来自世界 4 地的病人临床生理测量数据和诊断结果，用来研究如何从生理特征预测心脏病诊断结果。其中部分临床特征已被专家认定为是疾病相关特征，这就为验证变量选择方法提供了一个参照标准。实验结果表明，与其他方法相比，CE 方法选择出了最多的疾病相关特征，在预测性和可解释性上优势明显。部分对比结果见图1。

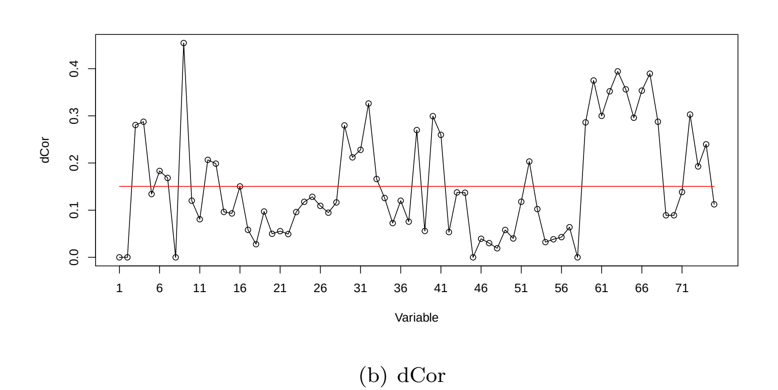
CE 为变量选择问题提供了统一的理论框架。它具有以下优点：

[4]实验代码：`https://github.com/majianthu/aps2020`

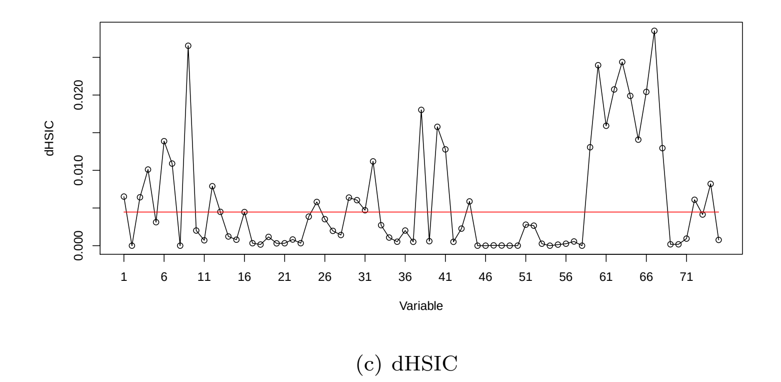
ChinaXiv:202105.00070v16



(a) CE



(b) dCor



(c) dHSIC

图 1: 三种统计独立性度量选择的变量.

ChinaXiv:202105.00070v16

- 模型无关，
- 数学理论坚实，
- 物理上可解释，
- 具有非参数估计算法，不做理论假设，
- 几乎不需要调参。

该方法做变量选择是模型无关的，这是与基于似然函数的方法相比，方法无需考虑模型及其复杂度等因素，具有明显的普适性优势。作为一种关联度量工具，CE 与其他度量工具相比定义更坚实，具有很多理想的独立性度量公理属性，因此也就具有了明显的理论优势。另外，熵是一种物理意义明确的数学概念，CE 可被认为是从自变量到目标变量的函数关系包含的信息量，因此很容易从物理上理解和解释得到的模型。在方法实现上，CE 的估计方法基于序数统计量，是非参数的，不做任何理论假设，充分发挥了其理论优势。同时，其估计方法具有良好的渐近稳定性，且几乎不需要调参，与 LASSO 等结果严重依赖超参数选择的方法形成了鲜明对比。总之，该方法具有理论和计算上的明显优势，将变量选择问题变成了一种科学，而不像 LASSO 等方法是一门艺术。

生存分析（Survival Analysis）是一类特殊的回归问题，其预测目标是事件发生时间（time-to-event），也即是未来某一事件发生所需要的时间。这类问题的特殊性还在于一种删失（Censoring）机制，用于当某一事件在观察期未发生时的处理。生存分析在医学、可靠性和社会科学等领域具有广泛的应用。建立生存分析模型也需要进行变量选择，用于筛选与事件发生时间相关的变量。马健 [76] 提出将 CE 的变量选择方法应用于此类问题，通过计算变量与事件发生时间之间的 CE 来选择变量。他将方法应用于两个公开的肺癌数据，与常用的随机生存森林（Random Survival Forest）和 Lasso-Cox 两种方法进行了对比，发现该方法能够在保证模型可解释性的同时获得更好的预测性能，验证了方法的优越性[5]。

3.4 因果发现

因果关系普遍存在于自然界当中，发现因果关系是各门科学的主要命题之一。从一组随机变量的时序观测中发现变量之间的因果关系，被称为因果发现（Causal Discovery）问题，是统计学中时间序列分析的经典问题。时序因果关系发现方法在不同学科领域都有重要应用价值。

如何度量因果关系是因果发现问题解决的基础。控制论学者维纳提出了一种因果关系的哲学概念，表述为因必须有助于改善果的预测 [162]。在此理念基础上，格兰杰提出了著名的格兰杰因果关系（Granger Causality：GC）检验 [163,

[5]实验代码：`https://github.com/majianthu/survival`

ChinaXiv:202105.00070v16

164]。GC 检验是经典的因果发现工具，但它只适用于高斯的情况。Schreiber [165] 定义了用于发现稳态时序包含的因果关系的传递熵（Transfer Entropy：TE）的概念。TE 是 GC 的非线性推广，等价于信息论的条件互信息（Conditional Mutual Information：CMI），本质上是检验条件独立性（Conditional Independence），是模型无关的，因此适用于任何情况的因果关系检验。TE 作为广泛采用的因果关系度量，较之其他经验式带有模型假设的传统因果关系推断方法更科学合理，具有更广泛的普适性。

CE 是统计独立性度量，而 TE 是条件独立性度量。我们证明了二者之间在数学上有着本质上的内在理论联系 [10]。通过并不复杂的数学变换，可以很容易证明，TE 可以表示为只包含 CE 的数学形式。这一数学表示形式为从 CE 估计 TE 提供了理论基础。

命题 1 *TE 可以表示为仅包含 CE 的数学形式. 从 X 到 Y 的 TE 的 CE 表示如下：*

$$TE_{X\to Y} = H_c(Y_{t+1}, Y_t) + H_c(X_t, Y_t) - H_c(Y_{t+1}, Y_t, X_t). \tag{7}$$

因为 TE 本质上是条件独立性关系 $Y_{t+1} \perp\!\!\!\perp X_t|Y_t$ 的度量，因此(7)也其实是给出了一种条件独立性的 CE 表示。

在过去的研究中，因果关系的估计往往是在一定的假设前提下进行，无假设前提的因果关系估计被很多研究者认为是不可能的。我们基于以上 TE 的 CE 表示形式，利用非参数的 CE 估计算法，提出了一种简单优雅、易于理解和实现的非参数 TE 估计方法 [10]。这样，不带任何假设条件的因果关系发现就成为了可能。此估计方法包含简单的两步[6]：

1. 利用非参数 CE 估计方法，估计式(7)中的 3 个 CE 子项；

2. 由 3 个 CE 估计值计算得到 TE。

为了验证提出的非参数 TE 估计方法，我们将该方法应用于大气污染问题中的因果发现，研究了北京地区气象因素和 PM2.5 之间的因果关系[7]。实验采用了 UCI 机器学习数据集仓库中的北京 PM2.5 数据 [166]，包含了北京地区 2010 年至 2014 年之间的每小时的连续气象观测数据和 PM2.5 观测数据。我们的分析选择其中一段无缺失值的连续时间数据记录，利用上述方法很容易就可以估计出气象因素对 1 至 24 小时后 PM2.5 浓度的影响程度。利用上述估计方法并不是无条件的，我们默认假设了时序是稳态的，也假设了时间段之间的马尔科夫性，也就是不相邻的时间段之间无关。对 24 小时内滞后因果关系的分析发现，温度、湿度、压力等气象因素对 PM2.5 的形成的因果关系是一个由迅速增加到缓慢增强的过程。

[6]此方法已在 R 和 Python 的 **copent** 包 [141] 中实现。

[7]实验代码：`https://github.com/majianthu/transferentropy`

ChinaXiv:202105.00070v16

同样在上述实验数据的基础上，我们将提出的 TE 估计方法与另外两种条件独立性度量进行了对比实验，估计从气象因素到 PM2.5 的因果关系 24 小时走势。这两种度量分别是基于核函数的条件独立性度量（Kernel-based Conditional Independence：KCI）[167] 和条件距离相关（Conditional Distance Correlation：CDC）[168]。论文通过将用 CE 估计 TE 与其它两种方法进行了对比，结果（见图2）显示 TE 的估计效果更好。

3.5 系统辨识

微分方程是描述动态系统的主要数学工具，在不同学科具有广泛的应用。从数据中学习微分方程是动态系统领域的一个重要问题，也称系统辨识或方程发现，近年来得到了大量的研究。

方程发现问题通常可以被当作一个回归问题来对待，即从数据学习一组从系统状态到状态微分的回归方程。给定一个一般的动态系统微分方程形式，如下：

$$\frac{dx_i}{dt} = f_i(\mathbf{x}, t), \tag{8}$$

其中 $x_i, i = 1, \ldots, N$ 表示系统状态变量，则方程发现问题就是从数据辨识 f_i。从数据辨识 f_i 需要确定该方程包含的未知自变量，一旦自变量确定则方程的对应关系就知道了，这是典型的变量选择问题。很多经典回归模型方法被应用到此问题，如高斯过程、基于稀疏性的方法（如 SINDy）、核函数方法和神经网络等。

马健 [15] 提出了一种基于 CE 的微分方程发现方法，将问题理解为变量选择问题，利用基于 CE 的变量选择方法解决了此方程发现问题。该方法包含了两个主要步骤：

1. 利用差分算子近似计算状态变量的微分；
2. 计算状态微分和状态变量之间的 CE，根据 CE 来选择方程的变量。

该方法中的差分算子可以由以下非参数方式计算得到：

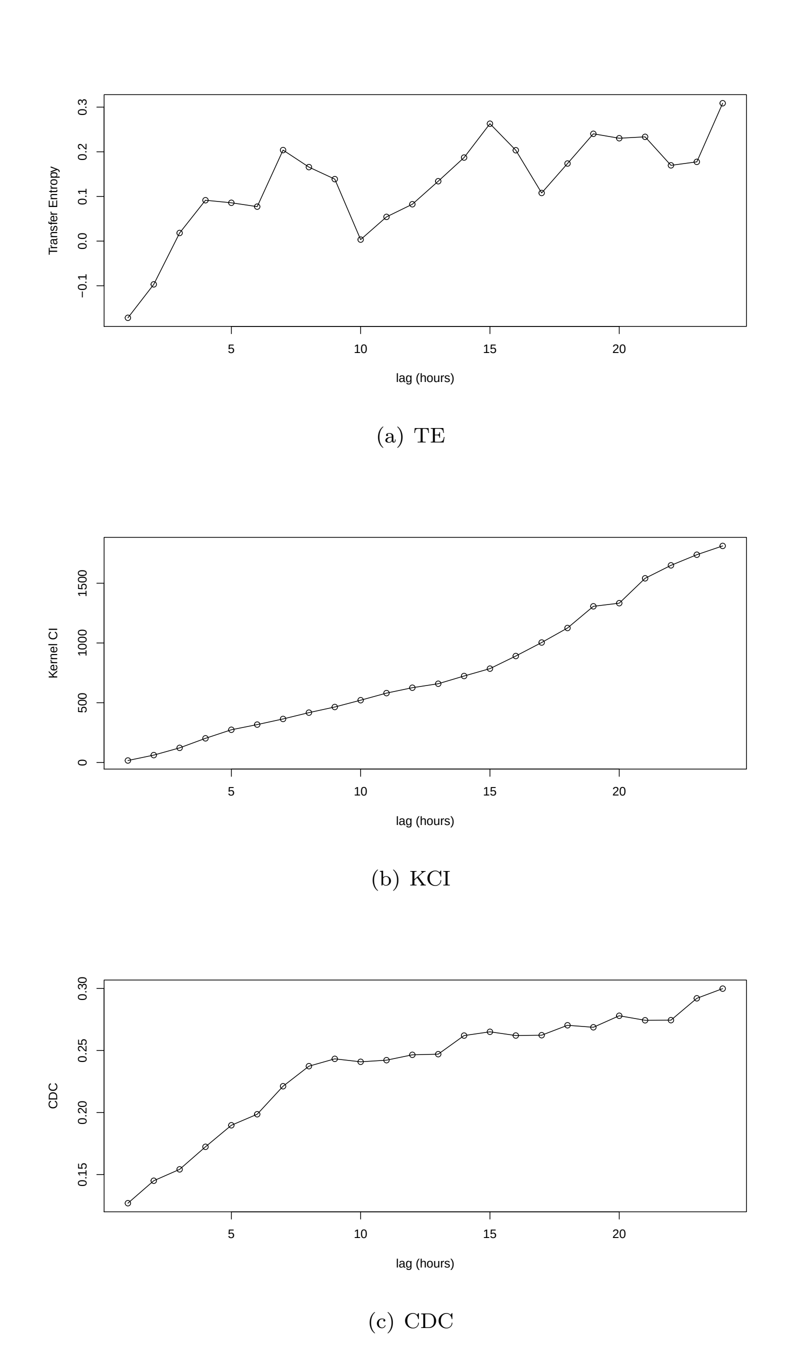
$$\frac{dx}{dt}|_{t=t_0} = \frac{x_{t_1} - x_{t_0}}{t_1 - t_0}. \tag{9}$$

而 CE 可以由非参数估计方法得到。因此，所提出的方法是非参数的，不做任何假设，适用于任何动态系统的辨识。

作者将方法应用于经典的 3 维 Lorenz 系统，系统中包含了由一阶和二阶的状态变量组成的 3 个方程，该方法成功地从仿真数据中辨识出了系统方程中状态变量和状态微分变量之间的关系，证明了方法的有效性[8]。

[8]实验代码：`https://github.com/majianthu/sysid`

ChinaXiv:202105.00070v16



(a) TE

(b) KCI

(c) CDC

图 2: 由三种因果关系度量估计的从压力到 PM2.5 的因果关系强度变化图.

ChinaXiv:202105.00070v16

3.6 时延估计

系统辨识（System Identification）是现代系统理论中研究系统特性的重要工具方法，主要是研究如何从观测数据来确定描述系统行为的模型及其参数。时延（Time Lag）是一种动态系统中普遍存在的特性，指一个变量作用于另一个变量需要的时间。由于物质、能量或信息的传输时间，时延存在于所有物理、社会和生物系统中的因果效应发生的时间先后关系上。因此，时延参数的估计是重要的理论问题，具有广泛的应用价值，比如可以用来分析交通系统中的拥堵传播、太阳活动对地球系统的影响、政策效应的分析等诸多问题。

传统的时延估计的主要方法是基于时序变量的自相关系数，但其具有线性假设，因此作用范围十分有限。另一种主要方法是时延互信息（Time-delayed MI），可以去除线性假设的限制，适用于具有非线性特性的时序变量。但这两种方法本质上都是对称关系的度量，而系统时延由于是因果关系的属性，因此是非对称的关系。TE 作为一种非对称的因果关系度量，量化了从因变量到果变量作用关系的信息量，因而更适用于估计因果时延特性。

动态系统的时延可以通过从时序数据估计 TE 统计量来进行估计，但传统的 TE 估计问题被认为十分困难。马健 [14] 提出利用上述基于 CE 的 TE 估计方法来解决时延估计问题，先估计时延窗口内的因变量到果变量的一组 TE 值，再将 TE 的最大值对应的时延作为时延参数的值。由于该 TE 估计器是非参数的，因此不对动态系统做任何假设，具有普适性。

作者仿真了四个具有不同动态特性的时延动态系统以验证方法的有效性，发现该方法可以准确地从系统的仿真数据中估计出相应的时延参数。作者又将方法应用于摩洛哥缔头万（Tétouan）城的电力负荷数据，分析五种天气因素对该城三个区域电力负荷影响的时延特征，发现了不同天气因素对负荷产生影响的时延长度，以及影响的每日变化特征[9]。

3.7 域自适应

域自适应（Domain Adaptation：DA）是一类常见的问题，是指训练模型的数据与应用模型的数据的概率分布不同时，需要让训练的模型适应分布的偏移的情况。DA 问题具有重要的现实意义。比如，将在一个医院采集的数据上训练好的模型应用到其他医院时，可能由于数据采集设备的不同导致采集的数据发生分布偏移，从而导致模型性能下降。同样的情况也会发生在其他领域（如社会学）的问题中，比如由于人群的社会属性的不同，由一个人群研究得到的模型结论在另一个人群上就会发生模型偏差。

基于 CE 的条件独立性度量作为一种基本的统计学工具，可以用于解决 DA 问题。马健 [11] 提出了一种从因果角度解决 DA 问题的方法。他假设自变量 X 到预测变量 Y 在不同域 D_i 上的关系是不变的，将数据分布迁移视为一个由外

[9]实验代码：`https://github.com/majianthu/timelag`

ChinaXiv:202105.00070v16

在条件变量 Z 在 D_i 上作用不同导致的结果，这样 DA 问题就转化为学习自变量 X、预测变量 Y 和外在变量 Z 之间统计关系的问题，需要发现 X,Y 之间不变的依赖关系，二者的依赖关系以外在变量为条件，即判断是否

$$X \not\perp\!\!\!\perp Y|Z. \tag{10}$$

这时，利用基于 CE 的条件独立性测试就能发现域迁移条件 Z 背后的 X 和 Y 之间不变的因果关系，从而很好地解决了 DA 问题。

作者设计了仿真实验验证了方法的有效性，并将方法成功应用于社会学的男女收入不平等的社会原因分析问题[10]。

3.8 正态性检验

正态分布是一类非常重要的概率分布函数，在所有概率函数中居于中心地位，且在实际问题中大量存在。正态性是很多统计模型和方法中的假设条件，因此在应用中检验分布正态性的方法成为了一个十分必要的工具。正态性检验（Normality Test）是一类检验分布正态性假设的假设检验方法，分为单变量和多变量两类。传统的正态性检验方法很多，比如基于矩、特征函数、熵或最优传输等概念工具的方法等。

CE 作为衡量变量间全阶次相关关系的度量工具，也可以用于检验二阶相关特性的正态性检验问题。我们都知道，根据最大熵原理，在二阶统计量相同的情况下，在所有分布中正态分布的熵最大。由于正态分布完全由其一阶和二阶统计量决定，因此正态分布的 CE 中包含的是二阶相关性对应的信息。可以很容易推导，在正态分布的条件下，CE 与二阶统计量具有等价关系。具体地，假设高斯随机向量 $\mathbf{X}_n$ 的协方差为 $V_{\mathbf{x}}$，则其 CE 与 $V_{\mathbf{x}}$ 之间存在如下关系：

$$H_c(\mathbf{x}_n) = \frac{1}{2}\log|V_{\mathbf{x}}|. \tag{11}$$

而在非正态分布中，相关关系不仅是二阶的，也有高阶的，因而 CE 包含的信息除了二阶相关关系对应的信息外，也有高阶相关关系对应的信息，且非高斯性越强，CE 中高阶相关的信息越多。同时，由于熵是全阶次的信息量，因此 CE 度量了包括二阶相关信息在内的全部阶次相关性信息。

马健 [12] 利用高斯分布的 CE 与二阶统计量之间等价关系提出了一种多变量正态性检验的方法，通过计算分布与同方差的高斯分布在 CE 统计量的差值来衡量联合分布的正态性，由此得到的正态性检验的统计量定义为

$$T_{ce} = H_c(\mathbf{x}) - H_c(\mathbf{x}_n), \tag{12}$$

其中，$\mathbf{x}_n$ 是与 $\mathbf{x}$ 具有相同方差的高斯随机向量。易知，当分布为高斯分布时 $T_{ce}=0$；且多元分布的非高斯性越强，T_{ce} 的数值越大。

[10]实验代码：`https://github.com/majianthu/cda`

ChinaXiv:202105.00070v16

马健同时给出了此统计量的估计方法，包括了十分简单的两部分：式(12)的第一项可以由 CE 的非参数估计方法得到，第二项即是式(11)的值，可先估计协方差矩阵 V_{x} 再解析计算得到。

作者设计了两组仿真实验，仿真了两类非高斯性的情况，并将此检验方法与 5 种经典的同类方法进行了对比，证明了此检验方法的有效性和对传统 5 种经典方法的优越性[11]。

3.9 双样本检验

双样本检验（Two-sample test）是统计学中另一类重要的假设检验方法，用于测试两组样本是否来自同一个分布函数。很多统计学的理论方法可以转化成双样本检验问题，如对称性测试就可以转化成检验对称变换的样本是否同分布的问题，又如变点检测（Change Point Detection）其实就是寻找一组双样本检测中样本间差异最大的点。同时，双样本检验又具有广泛的应用价值，比如可以检测临床治疗、政策实施等人为干预前后目标变量是否发生了变化等。常见的双样本检验方法很多，如双样本 T 检验，K-S 检验和基于核函数的检验等。但这些方法都有各自的不足之处，比如 T 检验需要正态分布假设，K-S 检验只能作用于单变量情况，而核函数方法需要超参数的调试等。

马健 [13] 提出了一种基于 CE 的双样本检验方法，思想是基于样本与检验标注之间的相关性程度来定义检验统计量。给定两组样本 $\mathbf{X}_0 = \{X_{01}, \cdots, X_{0m}\} \sim P_0$ 和 $\mathbf{X}_1 = \{X_{11}, \cdots, X_{1n}\} \sim P_1$，该检验的零假设为

$$H_0 : P_0 = P_1, \tag{13}$$

对立假设为

$$H_1 : P_0 \neq P_1. \tag{14}$$

定义 $\mathbf{X} = (\mathbf{X}_0, \mathbf{X}_1)$，以及两个检验标注变量 $Y_0 = (1_1, \cdots, 1_{m+n})$ 和 $Y_1 = (1_1, \cdots, 1_m, 2_1, \cdots, 2_n)$，则该检验的统计量可通过零假设和对立假设的 CE 差来定义：

$$T_{ce} = H_c(\mathbf{X}, Y_0) - H_c(\mathbf{X}, Y_1). \tag{15}$$

易知，当 H_0 为真时，T_{ce} 较小；而当 H_1 为真时，则较大。作者给出了基于 CE 非参数估计的统计量估计方法。因此，该检验方法是多变量非参数检验，且无需调参。作者在 3 组由正态分布和正态 Copula 仿真的数据上验证该方法的有效性，并将方法与基于 MI、核函数和 dCor 的三种多变量非参数检验方法进行了对比，发现该方法有效检验了仿真实验中的双样本假设，与同类方法相比具有同等或更好的检验性能[12]。

[11] 实验代码：`https://github.com/majianthu/mvnt`

[12] 实验代码：`https://github.com/majianthu/tst`

ChinaXiv:202105.00070v16

4 讨论

4.1 理论应用之间的联系

以上介绍的 CE 的前四个理论应用之间有着内在的联系。从理论基础上讲，它们都是基于 CE 对统计独立和条件独立的度量的理论框架，学习某种内在的统计关系，这是共同点。区别在于这四个应用研究的关系不同，以及关联结构的表示方式不同。关联发现问题主要关注成对变量之间的静态的统计相关，表示为相关矩阵的形式；结构学习则关注一组变量之间整体的关联结构，表示为图的形式；变量选择的目的是要建立一个多对一的关联结构，最终要表示为函数的形式；时序因果发现是动态系统中的因果关系，它也可以构建表示变量之间因果关系的有向图结构，也可以用来进行变量选择，构建时序预测的函数关系模型。

总之，利用 CE 度量统计独立和条件独立关系，可以估计随机变量之间的相关性和因果性关系强度，进而通过相关或因果关系发现表示成基本的矩阵形式，通过结构学习生成直观的无向或有向图的形式，或者通过变量选择构造具有预测能力的静态或动态时序的函数模型的形式。

4.2 相关性和因果性

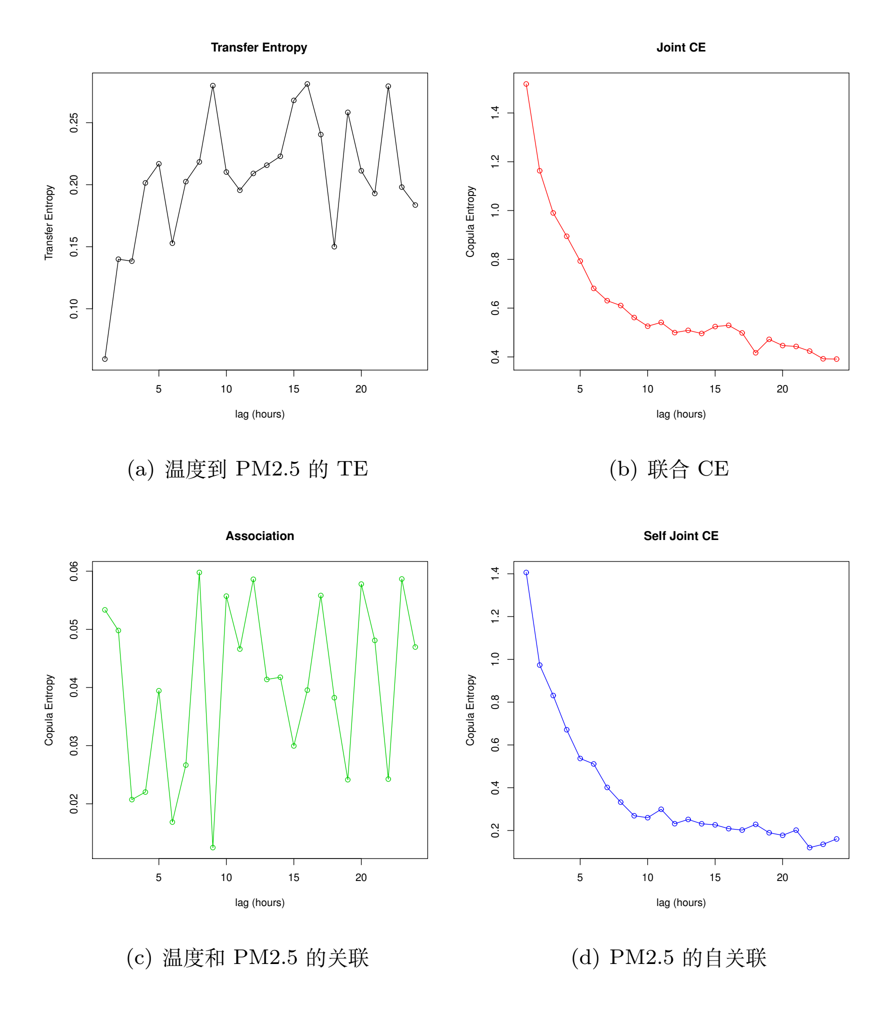
相关性和因果性是统计学中的两个基础性概念，对应于概率论中的统计独立和条件独立。统计独立和条件独立是两个不同的概念，但又有着内在的联系。我们通过 CE 的概念，给出二者之间的内在联系的理论框架，以及在此理论框架基础上的估计方法。

前者可以用 CE 来衡量。CE 是一个完美的衡量统计独立性/相关性的数学概念，具有很多数学家梦寐以求的独立性度量的公理属性。它等价于信息论中的 MI 概念。后者可以用 TE 来衡量。TE 等价于条件 MI。我们证明了 TE 可以用 CE 来表示。也就是说，条件独立可以通过统计独立来表示和计算。因此二者之间具有内在的理论联系。后者可以用 TE 来衡量。TE 等价于条件 MI。因此，二者之间具有内在的理论联系。

相关性不等于因果性，二者是不同的概念，但人们有时却很容易误把二者等同起来。举一个我们做的时序因果发现的研究 [10] 作为例子加以说明。 论文给出了一种利用 CE 来估计 TE 的算法，并采用了一个环境气象的数据来验证 TE 估计算法 [10]。数据是北京的 PM2.5 观测数据，以及同时观测到的北京地区气象数据。论文实验分析了气象因素（温度、露点、气压和风速等）对 PM2.5 浓度的因果强度，用从时序观测数据中估计的 TE 来衡量，发现了二者之间的因果关系变化规律。

这里要强调的是论文的讨论部分。我们讨论对比了时序相关性和时序因果性，发现即使是气象因素和 PM2.5 浓度之间相关性微弱的情况下，二者之间仍

ChinaXiv:202105.00070v16



(a) 温度到 PM2.5 的 TE

(b) 联合 CE

(c) 温度和 PM2.5 的关联

(d) PM2.5 的自关联

图 3: 对温度到 PM2.5 的 TE 变化的分解.

ChinaXiv:202105.00070v16

然有时滞因果关系。论文以温度因素为例（图3），对此做了说明。子图 (a) 和 (c) 分别对应 TE 和 CE，也就是因果性和相关性。我们可以发现，相关性强度几乎为 0, 而因果性强度依然很高。

我们认为，这一分析结果是由时序观测的对象系统的动态性造成的，气象因素对 PM2.5 浓度变化的影响不是即时的，而是由于大气系统的内部运动过程，有一个滞后的效应所致。此时，时序变量之间没有即时的相关关系，但存在时滞的因果关系。

4.3 三种理论框架的对比

TE 估计方法将统计独立性度量 CE 用于条件独立性的度量——TE 的表示和估计。从而，我们就提出了一个基于 CE 概念，能够度量独立性和条件独立性两个基本概念的理论框架。与此类似，核函数的方法 [154, 167] 和距离相关的

表 1: 三种统计独立性度量的对比.

	CE	**DC**	**HSIC**
定义	基于 Copula 函数	相关性的非线性扩展	核函数空间的相关性
多变量	是	distance multivariance	dHSIC
独立性测试	是	total distance multivariance	需要满足核函数条件
条件独立性测试	TE	CDC	KCI
不变性	单调变换不变	线性变换不变	无
Pearson 相关	高斯性假设下等价	高斯性假设下等价	未知
计算复杂度	$O(n^2)$	$O(n^4)$	$O(n^4)$

方法 [156,168] 也可以应用到这两个概念的度量问题上，也分别构成了类似的理论框架。但基于 CE 的理论框架更优越，理论上，CE 的定义更严格；计算上，基于 CE 的估计方法也更简单优雅，普遍适用，且计算量相对要小。

我们利用表1对比了三种统计独立度量概念，可以看到 CE 具有多方面的理论优势。比如，CE 天然的是一个多变量的度量，而其他二者需要通过扩展定义来满足多变量的情况；CE 具有单调变换不变性和在高斯条件下与相关系数等价等属性，而 DC 也具有类似的等价关系 [155]，HSIC 则未知。在计算成本上，CE 计算复杂度低，而其他二者则具有较高的计算复杂度。

在变量选择和因果发现两个理论应用中，我们利用真实数据对比三种框架中的相应方法。实验结果也表明了 CE 框架的（条件）独立性度量工具均优于其他两个框架中的相应的工具，能够更高效、准确地发现更多的相关或因果关系。

4.4 独立性和条件独立性度量的评估

独立性和条件独立性是概率统计领域的基本概念，具有基础性的重要地位。从统计学初期的皮尔逊相关系数开始，如何度量这两种统计学的概念就一直是本学科关注的焦点问题，有大量的度量方法根据不同的思想或原则被提出来。这其中，就包括上述的 CE 等三种理论框架的方法。

那么哪一种方法是最理想的度量呢？理论上，为了回答此问题，Rényi [169] 曾经提出了著名的独立性度量的公理系统，包括了 7 条公理。Schweizer 和 Wolff [170] 在提出他们基于 Copula 的度量时，对 Rényi 的公理系统又做了修正。

如何从实验的角度评估对比这些度量方法是一个重要的问题。马健 [171] 设计了一组仿真实验[13]，对现有的 16 种独立性度量和 16 种条件独立性度量方法分别进行了对比（度量方法及实现见表2），仿真实验考虑了线性/非线性、高斯性/非高斯性、单变量/多变量等多种不同角度的组合。同时，作者也在 3 组实际数据上对上述度量的性能进行了对比。结果表明，基于 CE 的独立性和条件独立性度量在所有情况中都表现了最好的性能，给出了最合理的度量估计值。

总之，从理论的角度来看，CE 具有比其他度量方法更坚实的数学基础和性能良好的通用非参数估计方法；从仿真和实际数据实验的结果来看，CE 在实验

[13]实验代码：`https://github.com/majianthu/eval`

ChinaXiv:202105.00070v16

表 2: 评估的独立性和条件独立性度量方法及其软件实现.

算法包	独立性度量	条件独立性度量	语言
copent	CE [5]	TE/CI [10]	R
stats	Ktau [172]		R
energy	dCor [155]		R
dHSIC	dHSIC [154]		R
HHG	HHG [158]		R
independence	Hoeff [159],BDtau [160]		R
Ball	Ball [161]		R
qad	QAD [173]		R
BET	BET [174]		R
MixedIndTests	Mixed [175]		R
subcopem2D	subcopula [176]		R
EDMeasure	MDM [177]	CMDM [178]	R
FOCI	CODEC [179]	CODEC [180]	R
NNS	NNS [181]		R
RCIT		RCoT [182]	R
cdcsis		CDC [168]	R
GeneralisedCovarianceMeasure		GCM [183]	R
weightedGCM		wGCM [184]	R
KPC		KPC [185]	R
ppcor		pcor	R
parCopCITest		pcop [186]	R
causallearn		KCI [167]	Python
pycit		CMI1 [187]	Python
knncmi		CMI2 [188]	Python
fcit		FCIT [189]	Python
CCIT		CCIT [190]	Python
pcit		PCIT [191]	Python

上展现了全面且最好的估计结果。因此，作者认为 CE 理论给出了最为理想的独立性和条件独立性度量。

5 实际应用

5.1 理论物理学

热力学是一门古老的理论物理学分支，在 19 世纪由克劳修斯、波尔兹曼和吉布斯等人建立，研究物理系统的宏观状态（如温度）与其微观状态之间的理论联系。熵和热力学第二定律是其最为核心的理论内容。香农的信息论就是受热力学的熵概念启发而建立的。一直以来，热力学和信息论之间的理论联系就是相关领域的重要话题之一。CE 是从信息论领域提出的数学概念，它的物理意义和解释一直未得到研究。马健 [16] 将 CE 理论应用于平衡态相关粒子系统中熵

ChinaXiv:202105.00070v16

的推导和计算，给出了 CE 的热力学解释，建立了热力学和信息论之间的又一理论联系。

5.2 理论化学

变构效应（Allostery）被认为“生命的第二秘密”，是普遍存在于几乎所有蛋白质的生命现象。它是指变构调节分子与蛋白质结合，诱导结合位点以外的远点发生变化的调节效应。最常见的变构系统模型是变构二状态模型，描述了变构过程的热力学循环。此类模型假设了受体活化是二状态过程，这与 NMR 实验揭示的多模态过程不相符合。深入理解配体诱导的受体活化的分子机制需要构建新的理论来理解配体结合点和激活点之间的热力学耦合关系。Cuendet 等 [17] 提出了一种新的理论，称为变构景观（Allostery Landscape），定义了热力学耦合函数来量化生物分子系统中的热力学耦合。他们指出新函数与 copula 密度函数和 CE 有密切联系，CE 定义了变构系统的信息传输属性，即配体结合点和激活点之间的信息传输。他们将新理论应用到丙氨酸二肽的 N 端和 C 端的热力学耦合分析中。

5.3 化学信息学

化学信息学是化学和信息学科的交叉学科，通过表征化学结构为数据，解决诸如分子设计、化学反应模拟和规划等问题。定量构效是该领域的前沿问题，研究分子结构与分子理化性质之间的定量关系，以指导具有指定特性的分子设计，应用广泛。分子理化特性可以理解为分子结构的某种对称变换不变性，而从数据学习得到这种不变性变换是分子设计的关键目标。Wieser 等 [18] 将对称变换学习问题转化为信息瓶颈（Information Bottleneck）问题，提出了一种对称变换信息瓶颈（Symmetry-Transformation Information Bottleneck：STIB）方法。该方法将分子表征表示为由两个部分组成的隐含表示，其中一个部分对应不变性表示，基于 MI（CE）的变换不变性，设计了问题模型的学习算法。作者将算法应用于包含 13.4 万有机分子的 QM9 数据库 [192]，使用其中具有固定化学计量（$C_7O_2H_{10}$）的 6095 个分子的子集，并将其对应的带隙能量和极性作为目标不变性属性。实验结果表明，STIB 方法给出了能够学习出表征分子属性、带隙能量和极性不变性的对称变换，验证了方法的有效性。

5.4 材料学

耐热型含能材料是指具有高能量和高热稳定性的特殊材料，可以在高温的环境下保持稳定性质，因此是国防、航空航天和地质勘探等重点领域的关键性材料，如宇航和高超音速武器的推进燃料、深井钻探的炸药等。但此类材料数量稀少且实验研究具有极高危险性，因此设计此类材料是材料学家们一直努力攻克的挑战性难题。“从头设计”含能材料需要经历“设计-筛选-评估”的流程，

ChinaXiv:202105.00070v16

其中采用机器学习的方法构建材料结构-性质预测模型对设计的分子性质进行预测是材料分子筛选的关键步骤。传统的含能分子性质预测模型构建过程只采用了与热稳定性线性相关的分子特征，没有考虑与含能材料热分解温度具有非线性关系的因素，如晶体结构和堆积方式等。田杰 [19] 提出了一种结合皮尔逊相关系数和 CE 的特征选择方法，从分子拓扑结构和量子化学计算特征中选择与热分解温度具有相关性的特征，并构建预测模型。其中，CE 方法的引入是为了筛选和热分解温度具有非线性关系的特征。他收集了 460 个含能化合物，并生成了包含 286 个特征的数据集，并应用该方法筛选得到了 87 个特征，再将筛选的特征做为随机森林和 SVM 等模型的输入以预测化合物的热分解温度，最终得到了较传统方法更好的预测效果，交叉验证实验的预测误差控制在了 28.5°C。他们将方法应用于自己设计的分子生成器生成的分子，最终筛选出 16 个具有良好热稳定潜能且爆轰能力很强的含能分子，验证了方法的实用价值。

5.5 水文学

洪水是主要自然灾害之一，洪水预报是降低洪水损失和管理洪水资源的重要手段。基于降水数据的降水量-径流量模型可以用来预报一段时间后的洪水。但是，水系统具有复杂性和非线性的特点，导致建立这样的模型时选择正确的模型输入十分困难。陈璐等 [20–22] 提出利用 CE 的方法来选择输入并建立神经网络预报模型。相比于传统的方法，基于 CE 的方法可以建立高维模型且对单个变量的边缘分布不做假设，同时由 CE 来估计降水量和径流量的数量关系的误差更小。陈璐等将方法应用于建立金沙江流域的洪水预报模型，结果显示利用 CE 选择输入的神经网络模型取得了最好的预测效果。Li 等 [23] 基于 CE 和机器学习方法研究了长江上游的月径流预报问题。他们利用 130 个全球环流指数、7 个气象因子和高场和寸滩两个水文站的月径流量数据，采用 CE 等 3 种变量选择方法和 5 种机器学习模型进行组合构建预测模型。结果表明，CE 和 LSTM 组合在高场站获得了最优预测性能，而随机森林和 CE 组合在寸滩站获得了满意性能。Mo 等 [24] 提出了一种长期径流预报模型框架，结合了 CE、LSTM 和 GARCH 三种方法，其中 CE 用于筛选与径流有关的预报因子。与传统方法相比，CE 更适合因子间具有交互关联的复杂情况。他们将方法应用于洪泽湖和骆马湖的径流预报研究，结果表明，与传统方法相比，该框架中的 CE 方法不仅成功辨别了因子间的交互效应，同时还量化了每个预报期内各个因子的贡献度，从而选出了与预报有关的关键驱动因子，最终该方法框架得到了较对比方法更准确、更稳定且更可靠的预报结果。

干旱是另一类重要的水文事件和影响重大的自然灾害之一。频发的干旱严重影响着我国的经济社会安全，特别是黄河流域的干旱威胁尤其严重，迫切需要开展流域干旱驱动和预测的研究。温云亮等 [25] 利用 CE 理论分析了河南省 1951–2014 年逐月气象数据，发现在众多驱动因子中，降水量、气温、水气压和相对湿度对该地区干旱发生的影响最大。Huang 和 Zhang [26] 利用 CE 方法

ChinaXiv:202105.00070v16

分析了兰州地区 1957–2010 年的气象数据，以构建该地区的干旱预测模型，发现该地区的风速、气温、水气压和相对湿度是与干旱最相关的气象因子。黄春艳 [27] 研究了黄河流域的气象、水文和干旱之间的关系，探讨了干旱的驱动机制，给出了气象干旱和水文干旱的概念，并提出利用 CE 方法探究二者之间的动态非线性响应关系，通过分析黄河流域不同区域水文站的气象和水文干旱指数，得到了水文干旱对气象干旱的滞后效应时间，为应对干旱事件提供了参考。Ni 等 [28] 利用 MI 和 CE 之间的等价关系，提出了基于 MI 的藤 Copula 结构选择方法，并应用于黄河流域干旱识别中特征变量建模问题和多水文站流量相关结构建模问题中。Kanthavel 等 [29] 利用 CE 和藤 copula 等理论工具，提出了一种综合干旱指数，整合了标准化降雨指数、干旱监测指数、标准化土壤湿度指数和标准化径流干旱指数等四种指数，可以更好地同时反映相关水文气象变量和不同类型的干旱。CE 理论被用来衡量新指数与原始指数之间的相关性。他们将该指数应用于印度中部的达布蒂（Tapti）河流域的单月和四个月尺度的干旱研究中，验证了该指数的有效性，并揭示了该地区干旱的时空分布特点。Mohammadi 等 [30] 利用基于 copula 和 CE 理论的三种相关性度量估计方法，在伊朗三座城市（扎黑丹、恩泽利和马什哈德）1950–2017 年的水文观测数据的基础上，分析了三地的干旱变量（干旱强度、时长和时间间隔）之间的依赖关系。

水文气象观测网络是获取水文信息的基础设施。如何设计并优化网络站点是一个综合性的科学和工程问题。一个基本的设计原则是观测站点之间尽量统计独立，这样才能最大程度的获取水文系统的信息。MI 是衡量统计独立性的主要工具，但是其计算是一个难题。Xu 等 [31,32] 提出了一个基于 CE 的多目标优化的水文观测网络设计方法，包括两步：1）基于 CE 的信息传输将观测站点分组；2）对每个分组选择最优的站点组合。基于 CE 的计算方法不仅能够处理水文变量的非高斯性，同时在计算性能上也更可靠、更有效率。作者将方法应用于黄河流域伊洛河水文观测网络和上海雨量观测网络的设计。结果显示，CE 的方法计算精度更高，且可以应用于高维的多变量估计情况。同样基于最少重叠信息的原则，Li 等 [33,34] 提出了一个由两个子目标构成的网络优化目标，其中一个子目标基于 CE 而设计，用于衡量冗余信息量。作者将此方法分别应用于汾河径流观测网、北京市区以及太湖盆地的雨量观测网的设计和优化，结果表明了方法可靠且有效。徐鹏程等 [35,36] 提出利用藤 Copula 来构建站点关系网络，再基于估计的藤 Copula 来计算站点间的 CE 值，在此基础上提出了结合 CE 和克里金指标的站点优化目标，利用滑动窗口法选择优化站点。他们基于淮河流域 1992-2018 年的日降水量观测数据，利用该方法对该流域 43 个雨量观测站点网络进行了优化，结果表明该方法得到的网络能够较传统类似方法得到的网络更有效地获取降水相关信息。杨惜岁 [37] 提出一个结合联合熵比、冗余度比和 NSE 效率系数的站网优化准则，并基于 CE 理论提出了新的 MI 计算方法，提高了计算的准确性。他将方法应用于美国查克托哈奇（Choctawhatchee）

ChinaXiv:202105.00070v16

河流域的 14 个水文站点，进行站点优化研究，最终得到了只包含 5 个站点的网络，提高了站网的监测效率。

分析河流的干流和支流之间的相关性对水利工程设计、洪水预防和风险防控十分重要。三峡大坝作为长江上游河段的大型水利工程，其一个重要功能就是洪水控制，研究该河段的主要河流相关性对工程设计和安全运行具有重要参考价值。Chen 和 Guo [38] 提出利用 CE 来计算河流相关性的强度，他们将方法应用于包含了 5 条主要干支流的长江上游河段，基于干支流 1951-2007 年间的洪水记录数据计算河流间的相关性。他们发现河流之间总的相关性并不高，这与该地区的气候特征相符；相关关系最强的是岷江和沱江，这是由于二者距离最近，且属于同一降水区域；金沙江和岷江、沱江之间具有一定的相关性，对三峡大坝的洪水控制构成了一定的威胁；金沙江、嘉陵江、岷江和沱江对长江盆地的洪水发生具有显著影响。

不同河流和区域的洪水事件叠加易于形成复合洪水事件，但不同洪水过程之间的空间关系很难利用现有相关性分析方法来准确地描述和评估。Wang 和 Shen [39] 提出了一个整合藤 copula 和相关性评估的方法框架，其中利用了 CE 理论从藤 Copula 来估计 MI、CMI 和 R 统计量等相关性强度。他们将方法用于评估长江上游已鉴别的 102 个复合洪水事件中两种极端径流序列变量（洪峰流量和洪水流量）之间的关系。结果表明，该框架的多维 R 藤 copula 模型能够更好地描绘复杂多样的水文相关关系，特别是藤结构表示了支流洪水汇入干流的顺序和水文站之间的空间位置关系；该框架估计的三种相关性强度比传统的相关性强度更好地反映了复杂时空水文系统的复合洪水事件中的非线性关系。

黄河水沙调控关系到黄河治理的策略制定，科学认知评估黄河的水沙通量变化特征是基础性的科学问题，对研判黄河泥沙情势具有重要意义。特别是近几十年来，受气候变化和人类活动的叠加影响，黄河水沙含量发生了显著变化，需要准确估计径流量和输沙量的分布变化情况。Copula 函数是分析这种分布的基本数学工具，但此类问题往往观测样本较少，难以准确估计 Copula 函数的参数。Qian 等 [40] 提出了一种基于 CE 和全相关（Total Correlation）关系的 Copula 参数估计方法，用于解决在样本较少的情况下的 Copula 参数估计问题。他们将方法应用于黄河西柳沟河流域 1960-2016 年间年度径流量和输沙量的数据的分析，该流域在 1999 年前后水沙关系发生了显著变化，但数据较少。分析结果发现，对于 1999 年前后的两个时段，新方法均得到比两种传统方法更准确的 Copula 参数估计，对数据的拟合更好。

流域分区是水文学研究的重要方法，根据水文相似性特征划分流域内相似性区域，可解决无水文观测地区的水文计算等难点问题。径流响应是重要的流域水文特征，根据流域水文站点观测之间的相似性做流域分区是一种基本的研究路径。传统的流域分区方法基于相关性评价，往往难以反映水文系统内在的复杂关系。刘磊等 [41] 提出采用基于 CE 的 R 统计量来衡量节点间的径流相似性，再在此基础上利用社团检测算法对流域进行分区。他们将方法应用于鄱阳

ChinaXiv:202105.00070v16

湖水系，利用该流域的水文站观测对流域进行了分区，并将方法与传统的 K 均值聚类方法进行了对比。结果表明，该方法能够有效捕捉流域内湖库对径流的调节作用，从而得到较传统方法更合理的流域分区。

多站点径流生成是随机水文学的主要问题之一，生成的流量信息对任何水资源管理都是必不可少的。在径流数据记录有限的情况下，生成多站点径流数据十分必要，需要设计相应的数据生成模型。Porto 等 [42] 提出了结合广义线性模型（GLM）和 Copula 函数的多站点年度径流生成模型，前者表示时序结构，后者为多站点的空间相关性建模。在评价模型性能时，作者采用了包括 CE 在内的多个统计描述性指标，其中 CE 用来衡量非线性的全关联。作者将该模型用于生成巴西的雅瓜里比（Jaguaribe）-大都市水库系统的多站径流时序数据，结果显示模型表现出了优于当前最好水平的性能，特别是在衡量多站相关性的 CE 指标上，较其他模型更接近于历史观测数据。

南水北调工程是当今世界最大的水利工程，承担着从长江的汉江流域丹江口水库向北方地区城市调水的战略任务。准确的入库径流预报是科学合理的供水调度的前提条件，能够使工程更充分高效地利用自然界的水资源。但传统方法构建的预报模型很难满足调水预报精度的要求，原因在于传统分析方法不能处理水文系统的非线性特性，导致了构建的入库径流预报模型不合理从而预测性能不高。黄朝君等 [43] 构建了丹江口水库的月入库径流预报模型，利用 CE 选择了一组气象水文因子作为模型的输入，得到的模型具有明显优于传统模型的预报性能。模型成功的原因在于采用 CE 选择的预报因子与中长期入库径流密切相关，印证了印度洋偶极子事件和南海副高活动与汉江流域夏季强降水之间的内在联系，符合自然界水文系统的运行规律。

5.6 气候学

气候变化是气候学研究的课题之一，它不仅体现在水文气候变量幅度上的变化，也体现在变量的季节和周期变化的分布上。这种变化会对降水和气温的强度和频率造成影响，导致极端天气（如洪水、干旱和热浪等）的增加。降水和气温的相关性会加剧联合极端天气的发生和强度。研究气候变化对降水和气温相关结构的影响是一个重要的问题。Hao 和 Singh [44] 利用 CE 度量工具研究了气候变化对这种相关结构的影响。研究采用了美国德克萨斯州达拉斯市沃斯堡（Fort Worth）在 1948-2010 年间的每日降水和气温数据，以每 5 年为期计算温度和降水之间的负 CE 值作为相关结构强度，发现该地区的温度和降水之间的相关结构强度（负 CE 值）从 1948-1980 年间的 0.18 下降到了 1948-2005 年间的 0.06，说明了气候变化对该地区水文气候变量之间关系造成了影响。

气候评估是科学应对气候变化的基础性工作，其目标是监测和分析全球和地区气候及其变化，特别关注于变化趋势和极端气候风险等。气候分类是指根据相似气候特征将地区分类，最常见的 Köppen 分类法采用的气候特征是温度模式和季节性降水。Condino [45] 提出了一种基于 Jensen-Shannon 距离的动态

ChinaXiv:202105.00070v16

分类算法，其中基于 JS 距离的分类准则采用了基于 CE 理论的表示方法并进行估计。他将方法应用于欧洲气候评估问题，根据 1951-2008 年间欧洲气象观测站每日温度和降水数据对欧洲 25 座主要城市的气候进行分类。结果表明，其提出的算法成功区分了分别属于欧洲南部和北部气候带的城市群，当进一步考虑南北气候过渡带时，算法也得到了与欧洲实际气候情况相符的合理的分类结果。

5.7 气象学

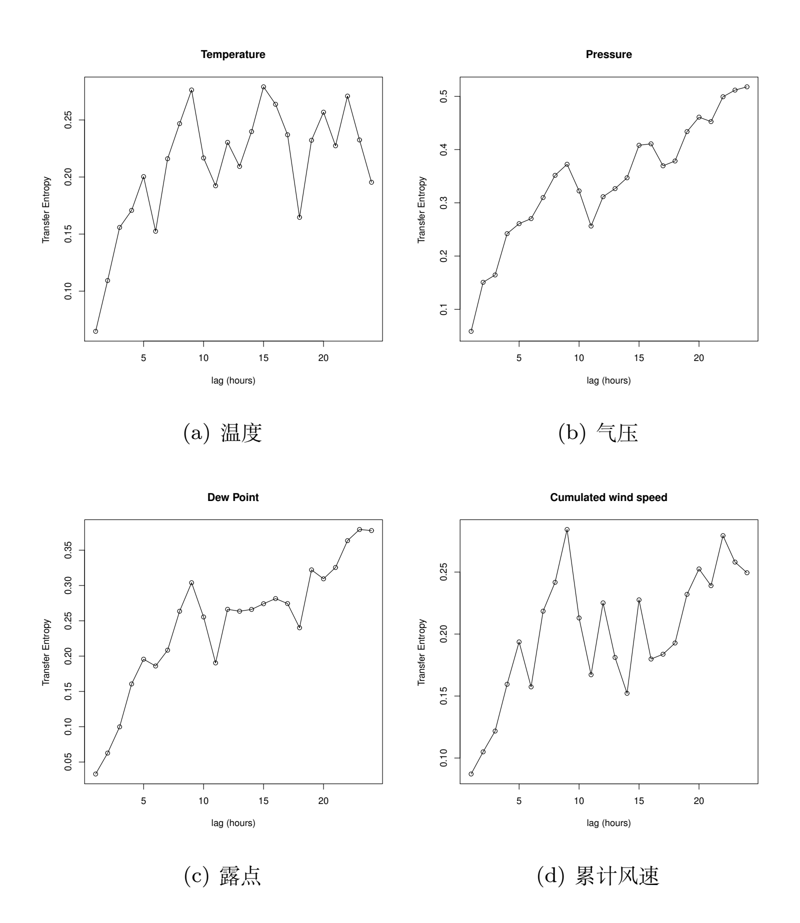
环境污染是现代社会的主要问题之一。从气象学的角度分析大气污染的成因，明晰其内在机理，有助于更好的理解污染问题，进而预测、干预和管理污染。理解大气系统中的因果关系是问题的关键。基于对气象因素和环境污染物的观测，可以利用统计学中的 TE 方法分析气象因素对环境污染的因果关系。马健 [10] 利用其提出的基于 CE 的 TE 估计方法（见3.4），分析了北京地区的气象和 PM2.5 连续观测数据 [166]，得到了四个气象因素对 PM2.5 浓度的 24 小时时滞内的因果强度变化图（见图4）。变化图显示，四种气象因素对 PM2.5 浓度的因果强度大致经历快速升高和缓慢增加两个阶段。作者还特别讨论和验证了该方法的平稳性假设和马尔科夫性假设在此中尺度数值分析问题上的适用性。论文所得到的因果变化图反映了大气系统运动的内在动态特征，增加了人们对 PM2.5 污染的气象成因的理解。同时，得到的时序因果关系也为整合气象因素，构建更优性能的污染预报模型提供了参考依据。（更多内容见3.4节）

有效的大气污染预测对于污染防控具有基础性作用，也利于保护居民健康。但当前的大气污染（如 PM2.5 浓度）预测在准确性和稳定性上还很难满足要求。开发性能更高的预测模型受到了广泛的关注。在综合考虑了传统方法的不足的基础上，Wang 等 [46] 提出了一种新的大气污染预测预警方法，使用了 CE 和多种机器学习模型的组合方法，CE 方法在其中被用来选择对 PM2.5 浓度波动有影响的因子，以用于构建最终模型。他们将开发的方法应用于上海和广州两地的实际大气污染预测预警系统，结果表明新方法能得到较其他对比方法更好的预测准确性和稳定性。Wu 等 [47] 提出了一种基于 CE 的 PM2.5 预测方法，利用 CE 计算气象因素与大气污染物浓度之间的相关性来选取模型输入特征，在基于 LSTM 和进化算法相结合的方法建立预测模型。该方法在北京地区 2016 年的历史数据上取得了良好的预测性能。

全球气候变暖导致我国华南地区的台风强度越来越强，强台风给该地区造成了严重的损失。根据台风灾害的观测数据预测灾情程度，是台风灾害的研判和应对的重要参考。但台风灾害影响因子较多，且与灾情之间具有非线性关系，给预测模型构建造成困难。陈燕璇等 [48] 基于 CE 等工具，提出了一种台风灾情预测模型构造方法。他们基于 1985-2014 年间登陆或影响广西的 44 个台风灾害数据，以及同期与致灾、承灾和防灾减灾相关的灾情统计数据，构建了 21 个灾害影响因子，再利用 CE 筛选与灾情指数最相关的因子，发现最大风速、最低气压、暴雨时长和暴雨极值与灾情指数最相关，能够客观地反映实际情况。实

ChinaXiv:202105.00070v16

ChinaXiv:202105.00070v16



(a) 温度

(b) 气压

(c) 露点

(d) 累计风速

图 4: 四种气象因素到 PM2.5 浓度的 24 小时时滞内因果强度变化图.

验也表明，利用 CE 筛选的因子构建的模型的预测精度要高于同类对比方法构建的模型，可为广西台风灾情预测提供参考。

5.8 环境学

大气污染是现代城市面临的主要环境问题之一，严重影响城市运行和居民生活。对大气污染扩散规律的分析是环境科学的重要问题，对环境监管部门更好地了解污染规律并有效应对具有基础性的指导作用。大量的城市污染气象观测网点生成的数据，有助于分析扩散规律进而对污染扩散作出预测。吴京鹏 [49] 提出了一种节点无特征网络链路预测算法，并将其应用于城市大气污染传播路径建模和预测问题。他将方法应用于兰州市辖区环境监测站网 2017 年的 PM2.5 观测数据，利用基于 CE 的 TE 算法构建了传播网络，再应用提出的网络链路预测算法预测污染传播路径。实验结果表明，该方法可以准确发现污染传播路径，为兰州城市污染治理策略制定提供了理论支撑。

氮氧化物（NOx）是火力发电厂排放的主要污染物之一，需要通过实施监测来严格管控其排放浓度。电厂一般采用 SCR 脱硝反应器的中和方法控制 NOx 排放浓度，但控制过程存在大延迟，无法做到精确控制，一般采用软测量模型预测与 SCR 控制器配合来达成控制目标。金秀章等 [50] 提出了一种 VMD-Bayes-Lasso 相结合的软测量算法框架，以预测 NOx 排放浓度。该方法框架首先利用 CE 筛选与 NOx 浓度相关的系统变量，以预测分解后的 NOx 浓度模态变量，再叠加得到最终预测结果，最后设计了基于 Lasso 算法的模型误差预测模型来校正预测结果。他们在宁夏某 660MW 燃煤电厂的数据上验证了该算法框架，得到了较对比方法更好的预测精度，其中，通过 CE 方法分析了系统变量内部以及和目标变量之间的相关度，达到了精简模型和提高预测精度的目的。

5.9 生态学

在生态学中，动物运动轨迹研究是一个重要的基本问题，可以揭示种群活动规律、种群间的竞争关系，以及种群和环境资源之间的互动等基本生态学过程。信息技术在生态领域的利用生成了大量的动物轨迹数据，对这些数据的分析需要合理的建模方法。环线数据（circular-linear data）是生态学中的一种常见的时序数据类型，描述了离散化的动物运动过程，包括运动方向和运动距离两个变量。此二变量之间通常是相关的，即直线运动时运动方向较小而运动距离较大，转向运动时运动方向较大而运动距离较小，同时运动方向变量的分布一般是对称的，因此通常采用角度对称的环线 copula 函数作为工具对此类数据进行建模，并利用基于 copula 的相关性度量来衡量二者之间的相关性。Hodel 和 Fleberg [51] 实现了环线 copula 的建模和分析的算法工具包 `Cylcop`，其中包含了基于 CE 的互信息估计算法作为相关性度量方法，用于分析动物轨迹数据。

ChinaXiv:202105.00070v16

5.10 动物形态学

动物形态学是动物学最古老的分支，研究动物体的形态和解剖结构以及其在发育和进化过程中的变化规律。作为动物学的基础学科，形态学的研究是动物分类的基础，比如鱼类的形态分类。由于鱼类的外形相似，对其种类进行鉴别往往会出现偏差，这就需要研究鱼类结构形态之间的相似性度量问题。Escolano 等 [52] 提出了一种图形形似度度量的估计方法，将图形转换为多维流形嵌入向量，再利用 CE 估计方法估计向量之间的 MI 作为图形相似度度量。他们将方法应用到 `GatorBait` 海洋鱼类图形数据库，该数据库包含了 30 个类别的 100 个鱼类外形三角网格图形。由于每类对应的是鱼类属而不是种，因此同一类别间具有形态差异，给分类造成困难。他们利用新度量方法对数据库中的鱼类图形进行分类，实验表明新的度量方法在数据集上得到了较传统方法更好的分类性能。

鲍是一类重要的海洋贝类，具有较高的营养价值和巨大的经济价值。鲍鱼的形态学研究是通过形态学变量的测量来研究其生长过程和种群分布等问题，对该类海洋资源的管理具有重要意义。Purkayastha 和 Song [53] 提出了一种新的因果关系度量概念，称为非对称 MI（AMI），用于判断变量之间因果预测性的方向，并基于 CE 理论给出其快速且鲁棒的估计方法。他们将 AMI 方法应用于 UCI 鲍鱼数据集，分析了鲍鱼的长度、直径、身高和体重等形态学参数的测量数据，明晰了鲍鱼生长过程中年龄与这些变量之间的因果关系规律。

5.11 农学

全球变暖导致的环境变化会直接影响粮食产量，从而加重世界粮食安全问题。水稻是最重要的谷物作物之一，占我国谷物产量的四成左右，对我国的粮食安全至关重要。研究气候变化如何影响水稻产量并给出对策是关系到我国粮食安全的重要问题。Zhang 等 [54] 利用作物模型和大气环流模式研究了气候变化对我国南方（江南和华南）两季稻生长和产量的影响及对策。研究采用了 DSSAT 作物模型中的模拟水稻生长和产量的 CERES-rice 子模块和 CMIP6 中的四种大气环流模型（GCMs），并利用 CE 和随机森林分析各个月份的气象因子和作物产量之间的非线性关系。他们利用每个 GCM 的 27 组数据驱动南方 54 个地点的水稻作物模型以得到最终产量，同时研究了播种日期的影响。研究发现，气象因子的上升趋势会提前水稻成熟期并降低产量；如果再考虑 CO_2 作用，早稻产量则会增加，而晚稻仍会减产；根据 CE 计算结果，两季稻产量和 CO_2 浓度的关系也是气象因子中最强的；提前早稻播种和延后晚稻播种可能会增加一定的产量。该研究的结论为政府和农民应对未来的气候变化指明了路径，为采用相应的适应性对策提供了重要参考。

ChinaXiv:202105.00070v16

5.12 认知神经学

认知神经学通过分析大脑活动的各种模态的观测数据，理解大脑作为信息处理器官，对外界刺激的表示、处理和通讯的机理。作为一个非线性的统计度量，MI 被认为是分析大脑信号间关联的理想统计工具。但由于 MI 的估计十分困难，使其难以得到广泛的应用。Ince 等 [56] 根据 MI 和 CE 之间的等价关系，提出了一种 MI 估计方法，称为高斯 Copula 互信息（Gaussian Copula Mutual Information：GCMI）。GCMI 方法利用了 CE 与边缘函数无关的性质，首先将每个变量的边缘函数转化为高斯函数，从而得到联合高斯分布，再根据所得高斯分布相关矩阵与 MI 的关系来计算 MI。该方法简单方便，且与分布无关。但由于从高斯分布数据计算 MI 是有偏差的，因此此方法还需要进行校正纠偏操作。Ince 等将 GCMI 与其他 MI 估计方法进行了对比，并将其应用于分析人脸检测任务的 EEG 数据 [55] 和听觉语音刺激任务的 MEG 数据 [57]。在人脸检测任务的实验中，GCMI 被用来计算图像内容与认知响应之间的关联强度，并成功选出认识响应敏感区域（图像中的眼睛部分）。在听觉刺激实验中，Ince 等研究了语音中的节奏特征对大脑听觉的节律同步的影响。通过对语音刺激的 EEG 响应数据的分析，作者发现了改变音节和词汇之间的停顿会导致听觉 delta 带同步的降低。在此实验中，GCMI 是数据分析的主要工具。

在 GCMI 算法的基础上，Combrisson 等 [58] 提出了基于信息论的群体层面分析大脑认知网络的方法，将非参数的排列操作与信息度量相结合，用于分析固定效果或随机效果模型，以适应多人间和多次任务间的变化。他们将方法应用于两个已有研究的数据：第一个研究分析人执行认知行为映射任务时的 MEG 数据中的高 Gamma 行为（High Gamma Activity），发现了任务相关的大脑网络，涉及多个运动区、体感区和视觉皮层区域等；第二个研究分析奖惩学习任务的前脑岛（anterior Insula）SEEG 数据，发现了奖惩任务的响应时延，以及奖和惩响应的显著差别。汪方毅等 [59] 提出了一种老年人认知水平分类方法，首先采用 GCMI 构建脑认知网络，再利用 GCMI 进行特征选择，最后利用 SVM 从选择的脑网络连接进行认知水平分类。他们将方法应用于 98 名葡萄牙老人的静息态 fMRI 数据，发现提出的方法能够捕捉数据中脑区间的非线性关系，并能够最终得到较同类方法更高的分类准确率。

语音理解是人脑的主要认知功能，研究人脑的神经活动对语音信息的编码和解析是认知神经学的重要问题。语音包络（speech envelope）包含了语音信号中的低频时序信息，研究表明其可以解释大部分神经响应的变化过程，语音包络跟踪就是通过脑电图等手段研究语音包络及其神经响应之间关系的问题。由于大脑的非线性特征，常用的线性模型不能很好的表示这种关系。MI 作为非线性关系度量工具，被认为能够捕捉语音包络和神经响应之间的非线性关系。De Clercq 等 [60] 利用根据 CE 理论构建的 GCMI 工具，基于两组故事讲述语音和相应采集的 EEG 数据，对比了线性模型和 MI 分析对大脑非线性成分的刻画能力。实验结果表明 MI 分析检测到了线性模型以外的显著的非线性成分，证明

ChinaXiv:202105.00070v16

了 GCMI 是比线性模型更适合于研究神经包络跟踪问题的工具。作者也实验验证了与传统的 MI 估计方法相比，基于 CE 原理的 GCMI 方法具有鲁棒、无偏和适合多变量分析等诸多优点。

建立神经信号之间的因果关系对理解脑连接至关重要，因果关系连接反映了在脑认知过程中脑网络内部不同区域之间的信息传输方向，刻画了大脑认知过程的脑区之间动态关系特征。相比于传统的格兰杰因果检验，无模型假设的 TE 更适合此类因果分析任务。Redondo 等 [61] 基于 CE 理论提出了一种新的 TE 概念，称为 STE（Spectral Transfer Entropy），用于计算频域滤波后的时域信号之间的 TE。与直接在原始信号上计算 TE 相比，在特定频域上计算的 STE 更具有神经学意义的可解释性。他们将方法应用于注意缺陷多动障碍（ADHD）患者 EEG 信号的分析，利用 STE 构建因果关系脑连接网络，发现了 ADHD 患者与健康人之间与注意力相关的脑连接网络连接的不同。实验结果表明，健康人在与注意力和受控记忆存取相关的 θ 和 α 频段表现出明显的因果联系，而 ADHD 患者的脑网络连接则主要在 δ 振荡上，可解释为与注意力缺陷有关。

5.13 运动神经学

肌肉协同（Muscle Synergy）是运动的基础，指人完成各种动作时肌肉组合之间时空上的动作协同。人体的运动控制系统是一个具有冗余自由度的系统，一般认为神经系统通过运动基元的组合协同策略来完成一个动作。运动控制研究的一个重要基本问题是鉴别运动控制中简化的基本肌肉协同策略。通过分解运动过程的肌电（Electromyographic：EMG）信号数据理解运动控制潜在的基本协同机理是基本研究手段，但如何处理信号中的非线性是主要的难题之一，基于 CE 的 MI 估计是处理此难题的有力工具。Wu 等 [62,63] 将多元变分模态分解与基于 CE 的 MI 相结合，构建了肌肉耦合网络模型，基于表面 EMG 数据分析了健康人伸手运动过程中上肢肌肉间的时空协同，成功刻画了肌肉耦合关系强度。Reilly 和 Delis [64] 提出利用基于 CE 的 GCMI 来度量 EMG 信号之间的时空关联关系，再利用矩阵分解的降维方法来发现 EMG 信号时空关联中的基本的肌肉协同模式。他们采集了人执行点到点动作运动的 EMG 数据，将方法应用于数据，得到了有生理学意义的肌肉协同时空模式。Zhu 等 [65] 提出了基于 CE 的表示 TE，再利用 R 藤 copula 估计 CE 进而估计 TE。他们将该方法应用于上肢肌肉间耦合网络的研究，基于疲劳/非疲劳状态下上肢肌肉运动的 sEMG 数据构建了肌肉耦合网络，发现疲劳状态下的肌肉群间耦合关系较非疲劳状态逐渐加深。金国美等 [66] 提出利用小波分析和 CE 估计相结合的方法，分析健康人自主运动下的肌肉疲劳状态的 sEMG 信号数据，发现在肘关节屈曲运动中，肌间耦合强度在 Beta 与 gamma 频段最为显著，协同肌肉对耦合强度比拮抗肌肉对耦合强度大；疲劳后的耦合强度相对于疲劳前有所增强。

ChinaXiv:202105.00070v16

5.14 计算神经学

计算神经学是利用计算理论和方法来研究和理解神经系统的功能和机理的学科，研究如何描述生物神经元对信号刺激的个体和群体响应等问题。神经可塑性（neural plasticity）是指神经元网络对外界刺激的适应性结构变化，构建可塑性理论模型是计算神经学关注的主要问题之一。Leugering 和 Pipa [67] 基于 Copula 理论提出了一个神经元群体可塑性的理论框架，构建了一种自适应网络模型，可以在未知模型输入变化的情况下保持模型输出的不变性，CE 在该框架中用于度量神经元群的统计特性，衡量输入输出之间的信息量。神经元之间的信息传输分析是计算神经学的另一个重要问题。分析计算神经元之间的信息传输关系需要涉及多个神经元之间的 MI 的分解。部分信息分解（Partial Information Decomposition）就是将 MI 分解为协同（Synergy）、冗余（Redundancy）和独特信息（Unique Information）三个部分的理论。基于 CE 理论和方法，Pakman 等 [68] 提出了一种估计独特信息的方法，并应用于分析多个神经元模型的信息处理。

5.15 心理学

大脑是一个分布式的网络系统。它不仅控制身体，改变内部生理状态，也影响多个高级过程。同时，内脏信息也时刻受到大脑的监控，也就意味着内脏过程也会反映到皮层活动中。内脏事件相关的大脑活动研究是一个重要的话题。植物神经系统中的过程之间相互关联，而信息论则提供了研究它们之间关系的工具。Ravijts [69] 研究了四种情绪刺激特征（效价、唤醒、支配和喜欢）下心跳诱发脑电位（HEP）的时间交互近似估计问题。他采用了用于情绪分析的生理信号 DEAP 数据集，利用基于 CE 的 GCMI 方法估计了 MI、协同和冗余等统计量，用于度量不同情绪刺激下 HEP 上的时间交互。实验发现了支配和喜欢情绪刺激下 HEP 上的时间交互现象，第一次揭示了情绪感知调制的 HEP 的时序特性。

5.16 系统生物学

系统生物学的一个主要任务是通过生化运动学模型，研究调控、信号传导和代谢过程之间的交互。建立这样的模型需要选择合适的模型输入变量，MI 是变量选择的工具之一。但常用的 kNN 的 MI 估计常常是有偏差的，需要进行修正。Charzyńska 和 Gambin [70] 提出了偏差校正方法，并发现当利用 MI 和 CE 之间的关系估计 MI 时，校正效果显著。作者将方法应用于受到广泛研究的 p53 蛋白和 Mdm2 连接酶之间的负反馈环路问题模型上，结果显示此方法能够比传统的本地敏感性分析方法得出更准确地反映系统行为的模型输入输出关系的分析结果。

系统生物学对分子生物学数据分析的主要目的之一是建立复杂生物现象的

ChinaXiv:202105.00070v16

网络和动态机制，以分析生命组织的功能和行为。MI 在构建基因通路网络的过程中发挥基础性作用。Farhangmehr 等 [71] 首次提出在网络构建中利用 CE 来估计 MI。他们将方法应用于酵母细胞周期数据，将分析得到的动态网络与京都基因组学百科数据库进行了对照。实验结果显示，利用 CE 来估计 MI 提高了计算效率。

5.17 生物信息学

生物信息学（Bioinformatics）是通过算法分析基因数据（包括基因表达谱数据）来研究生命和疾病机理的新兴学科。基因表达谱是利用 DNA 微阵列技术在基因分子层面观察某一生命组织动态得到的数据，从而能够在基因组水平上反映生命系统的各种现象和机理。Wieczorek 和 Roth [72] 提出了一种研究时间序列数据之间相互作用的分析方法，称为因果压缩（Causal Compression）。与传统的分析全时间序列之间的因果关系不同，该方法研究了基于定向信息（Directed Information）分解的时间序列间相互因果作用的稀疏表达，并据此给出了时序因果分割和因果二分图发现两类问题的解法。基于 CE 与 MI 之间的等价性，作者证明了该方法只与数据分布的 Copula 密度函数有关，并据此设计了求解方法。作者将该方法应用于 NCBI 数据库中的人类 C 型肝炎病毒感染数据（NCBI/GEO 查询号：GSE7123），研究了接受了聚乙二醇干扰素和利巴韦林治疗的重组丙型肝炎病毒核心蛋白基因型 1 感染的基因表达谱时序数据，关注了在干扰素信号传导中具有重要交互角色的两个基因：转录子 `STAT1` 和干扰素诱导抗病毒基因 `IFIT3`，分别生成了二者在有效救治和无效救治病人内相互作用的不同。研究发现，根据分析结果，干扰素疗法消除了大多数有效救治病人体内两种基因之间的关联，而无效救治病人体内的关联则不受影响。同时，分析表明两种病人救治前后二者之间均存在因果交互作用，但对于有效救治病人，早期的 `IFIT3` 对后期的 `STAT1` 的影响更显著，这与已有研究结论相符合。

很多疾病的发生与基因结构变异有关。拷贝数变异（Copy Number Variations：CNVs）指长度大于 1kb 的 DNA 片段的变异，在人类基因组中大量存在。作为重要的基因变异，CNVs 包含了大量 DNA 序列、疾病点和功能单元，能为疾病研究提供线索。研究表明，多种癌症的形成和发展与不同的 CNVs 有关。因此，发现不同基因的 CNVs 与不同癌症之间的关系有助于研究癌症病因和诊断方法。从大量的 CNVs 的基因特征中选择出与癌症相关的特征是生物信息学的一个重要问题。Wu 和 Li [73,74] 提出了一种基因选择方法，称为相关冗余和交互分析（Correlation Redundancy and Interaction Analysis：CRIA）方法，根据 CNVs 选择与癌症有关的基因，以用于癌症分类。CRIA 方法利用了 CE 的多变量相关性特性，设计了基因特征交互强度度量，用于筛选与癌症类型相关性强的基因。他们将该方法应用于 cBioPortal 的癌症基因组数据，利用了其中的 6 种癌症数据，选择出了 200 个与癌症有关的基因。为了验证算法的有效性，他们基于亚利桑那州立大学的数据将方法与其他 8 种基因选择算法进行

ChinaXiv:202105.00070v16

了对比，结果显示 CRIA 方法选择的基因能够更准确地预测癌症类型。

5.18 临床诊断学

心脏病是最常见的临床疾病之一。医生已经积累了丰富的心脏病临床诊断经验，可以通过各种生理测量结果作出诊断决策。在此经验基础上开发智能临床诊断模型是业界长期追求的目标，开发此类模型的关键在于选择一组生理测量变量来构建预测诊断模型。基于著名的 UCI 心脏病数据集 [144]，马健 [9] 提出采用 CE 作为变量选择方法，用以选择一组生理变量构建诊断模型。该数据集包含了来自世界四地真实的临床心脏病生理测量和诊断数据，其中 13 个生理测量变量被医学专家认定为是临床相关的。实验结果表明，CE 方法选择出了 13 个临床医生认定变量中的 11 个变量，是对比方法中最多的，从而得到了最好的预测准确率。同时，CE 方法还发现了认定变量以外其他与诊断相关的变量，为临床进一步检验提供了新的参考。（更多内容见3.3）

糖尿病是另一种常见临床疾病。对糖尿病人的病情管理与临床诊治结果（发病率和致死率）密切相关，因此建立严格的糖尿病患者住院管理流程对其安全十分重要，这就需要对病情管理标准进行分析研究。为了评估住院患者的救治效果，美国业界建立了健康事实（Health Facts）数据集 [193]，包含了 130 所美国医院和救治网络的糖尿病患者的数据。基于该数据集 1999 至 2008 年的 10 年间 101,721 名住院患者的数据，Mesiar 和 Sheikhi [75] 利用 CE 变量选择方法建立预测模型，用于从其他 49 个变量预测“是否已用药”变量，取得了良好的预测效果，在仅选择使用 20 个变量的情况下就获得了 97.2% 的准确率，增进了对用药相关变量的认识，构建了合理用药评价模型。

癌症预后是指基于癌症病情的临床表现和诊断结果，对病情的未来发展进行评估，以帮助进一步的临床决策。临床评估考虑的预后因子在评估中至关重要，但又常常数量众多，需要进行分析选择。比如，肺癌的预后因子就多达百种。预后模型是在预后因子的基础上建立的病人风险预测模型，是癌症治疗中重要的临床工具。马健 [76] 提出了一种基于 CE 的生存分析变量选择方法，并将其应用于预后因子的选择问题，以建立预测病人生存时间的预后模型。他基于两个公开的肺癌数据验证了该方法，发现其能选择符合临床标准的预后因子，并获得较同类方法更好的预测模型，在保证模型可解释性的同时具有更好的预测性能。

白内障是眼科的常见疾病，是导致患者失明的最主要病因。白内障超声乳化手术（Phacoemulsification）是世界各国治疗白内障的首选手术治疗方式。尽管该手术已十分成熟，但临床仍然可能会导致术后角膜水肿等并发症，从而影响视力恢复并造成患者不适。构建基于风险因素的角膜水肿风险预测模型在临床十分必要。Luo 等 [77] 提出利用 CE 方法构建术后角膜水肿风险预测模型，将方法应用于临床 178 名患者的数据，从数据的 17 个变量中筛选预测变量，最终将临床预测模型使用的四种变量（糖尿病、最佳矫正视力、晶状体厚度和累积耗

ChinaXiv:202105.00070v16

散能量）减少为两种（最佳矫正视力和累积耗散能量），且不影响预测精度。结果分析表明，利用 CE 得到的预测模型具有临床应用价值，可以在保证预测性能的情况下减少预测需要收集的临床信息。

主动脉瓣反流（Aortic Regurgitation）是一种常见的心脏瓣膜疾病，主要症状是在心脏舒张期，血液从主动脉回流到左心室。主动脉瓣膜置换手术是主动脉瓣反流的传统治疗方式之一。左心室射血分数（Left Ventricle Ejection Fraction：LVEF）是一项衡量心脏功能的重要指标，研究其在手术前后的改善关系可以为瓣膜置换手术时机选择和效果预测提供参考证据。Sunoj 和 Nair [78] 利用 survival copula 扩展了 CE 概念，提出了一种称为 Survival Copula Entropy（SCE）的新概念，用于衡量生存函数相关变量之间的依赖关系。他们将 SCE 应用于主动脉瓣置换手术临床数据，发现了手术前后 LVEF 之间的正相关关系。

5.19 老年医学

阿尔兹海默病（Alzheimer's disease，也称痴呆症）是老年人面对的主要神经退行性疾病之一，临床表现为认知能力的过度衰退等。早期筛查和诊断可以帮助痴呆症患者和家庭及早干预并管理病情发展，可以有效提高病人生活质量，降低家庭和社会成本和负担。简易精神状态量表（Mini-Mental State Examination: MMSE）是临床广泛采用的认知能力筛查工具之一。马健 [79] 通过利用 CE 分析了手指扣击运动（finger tapping）的特征和 MMSE 之间的关联强度，发现一组与 MMSE 相关联的特征，包括扣击频率（或扣击次数或扣击平均时间间隔）等。在此关联关系的基础上，他们构建了从手指扣击特征到 MMSE 的预测模型，取得了良好的预测效果。此预测模型有望用于痴呆症等疾病的认知能力筛查工作中。

帕金森病（Parkinson's disease：PD）是另一种常见的神经退行性疾病，临床表现为动作迟缓和运动功能障碍等症状。重复经颅磁刺激（repititive transcranial magnetic stimulation: rTMS）是利用脉冲磁场作用于中枢神经系统，以改善生理功能的临床治疗技术，广泛应用于神经、精神类疾病的治疗，并在近年应用于 PD 康复治疗的研究中，以期缓解患者症状并改善运动功能。李润泽等 [80] 研究了 rTMS 对 PD 患者运动症状辅助治疗的神经调控机制，利用基于 CE 的 GCMI 等方法分析了 rTMS 治疗前后的 EEG 数据，构建了脑功能网络连接矩阵并得到 3 种网络特征参数。实验结果表明 rTMS 主要改变 PD 患者的 beta 和 gamma 振荡，其中运动皮层的相应变化可能与运动功能改善有关。

跌倒是老年人面对的重大健康风险之一，需要科学管理和及早干预。跌倒预测是管理跌倒风险的重要手段之一。起立行走试验（Timed Up and Go: TUG）是一种主要的跌倒风险评估工具。马健 [81] 提出了一种结合视频分析和机器学习技术的跌倒风险预测方法。该方法首先从老年人进行 TUG 测试的视频中分析出人体 3D 姿态信息，再由一段时间的姿态信息序列计算出一组步态特征，通过利用 CE 分析步态特征和跌倒风险指数之间的关联关系，选择出一组与风险

ChinaXiv:202105.00070v16

关联的步态特征（包括步幅、步态速度和步态速度的方差等），最后用此特征作为输入构建跌倒风险的预测模型。该方法在真实数据上的实验显示了良好的预测效果。此分析结果也表明了步态特征反映的行动能力与跌倒风险之间的内在联系，使得模型具有临床意义的可解释性。

在以上两个研究的基础上，马健 [82] 还利用 CE 对手指扣击运动特征数据和步态特征数据进行了联合分析，发现了某些手指运动特征与跌倒风险之间具有一定的关联性。这一发现为首次发现，揭示了衰老过程中认知能力和行动能力之间的关联，提供了科学实验证据，加深了对衰老的生理特征的认识和理解。

5.20 精神病学

抑郁症是一种常见的情绪相关的心理精神障碍，全世界约有 3.5 亿名患者为此病所困扰，对其进行研究对人类健康具有重要意义。脑电图（EEG）是一种非侵入式的大脑活动电信号测量手段，广泛应用于大脑疾病的研究中。脑功能网络是在 EEG 信号基础上构建的反映大脑活动的功能性指标，可采用 MI、相干性等多种方法构建此类网络。张婷婷等 [83,84] 提出基于相干性虚部（Imaginary part of Coherency）构建的脑网络连通性指标来研究抑郁症患者识别问题。他们利用 CE、Relief 过滤等特征选择方法对脑电网络连通特征进行选取，发现利用 CE 和 Relief 过滤联合得到的相干性在线反馈指标特征集合能够有效区分抑郁症患者和健康人群。

5.21 公共卫生学

流行病是公共卫生学的重要话题，流行病患者的及时诊断对控制流行病的传播至关重要。感染了流行病毒的病人往往伴有发热等症状，很难与正常的发热病人进行区分。目前正在流行的新型冠状病毒患者就具有这样的发热症状，基于临床数据开发能够区分病毒感染者和正常流感病人的技术成为一个紧迫的问题。然而，相关的症状有 10 几种，如何选择合适的变量集合成为研究成功的关键。Mesiar 和 Sheikhi [75] 基于 CE 变量选择方法，利用真实的临床数据，分析了新冠患者诊断相关的 19 种症状变量，发现年龄、疲劳和恶心呕吐是最重要的诊断变量，可以使诊断达到 85% 的诊断准确率，如果将诊断变量增加到 15 个，准确率可以提高到 91.4%。

高血压是全球首要致死病因，对人群健康构成严重威胁。全基因组关联研究表明多个基因与高血压密切相关。已有多个研究报道 I 型细胞膜钙离子转运酶基因（ATP2B1）与收缩压和舒张压相关联。该基因有 21 个 CpG 位点。研究该基因及其 CpG 位点与高血压的关系是一个新的重要问题。Purkayastha 和 Song [53] 提出了一种新的非对称可预测性概念，称为非对称 MI（AMI），并利用 CE 理论给出了其估计方法。他们将该方法应用于 ELEMENT 数据集，分析 525 个年龄在 10-18 岁之间的儿童的数据，发现 ATP2B1 与舒张压相关联，证

ChinaXiv:202105.00070v16

实了已有的发现；同时发现该基因的 CpG 位点 CG17564205 与舒张压相关联，且根据 AMI 判断，舒张压对该位点具有预测性，这一新发现表明血压可以改变位点。

5.22 经济学

经济政策的评估需要定量分析，定量分析方法可以科学、客观地评估政策效果。Shan 和 Liu [85, 86] 提出了一种可以定量分析政策组合效果的决策树构建方法，CE 被用来度量非线性相关关系并构建决策树，方法的思想是利用基于 CE 定义的信息增益来构建用以区别不同政策对象群体的政策决策树，由树的叶子节点来表示不同政策组合对应的群体划分。他们将该方法应用于发展经济学领域，评估我国的减贫政策效果，研究分析了 2018 年由政府开展的贫困家庭状况普查的问卷调查数据中四川省的数据。分析发现，就业政策、新收入来源和是否有抵押贷款是影响家庭收入的主要政策因素，并揭示了这些政策组合对应的不同目标贫困群体收入结构的不同特征。该方法在无历史数据的情况下，评估验证了减贫政策的有效性，并发现了更加有效的政策组合方案。

经济学的核心目的是发现因果关系。传统的经济学依靠推理建模以及基于此的实验设计。因果发现是从数据中发现因果关系的方法，将其与经济学理论模型相结合是设计经济学实验的新路径。Bossemeyer [87] 基于 CE 和 MI 的关系提出了一种条件独立性测试算法，并将其应用于因果结构发现的 PC 算法中。作者利用新 PC 算法研究了经济学中的议价理论，研究讨价还价行为中互惠关系的作用，以及响应时间在这个过程中的作用。作者将算法应用于 eBay 的 Best Offer 平台数据，发现交易双方让价行为之间存在关联，印证了互惠理论；同时，发现了对手还价响应时间对下一次要价存在因果效应。

产业链是指产业部门之间基于经济关系形成的链条式关联关系形态。产业链基于资源要素分配和专业化分工等多种因素构成上下游关系，来进行价值互换，上游企业向下游提供产品和服务，下游向上游反馈信息，由此形成关联互动关系。产业链各环节之间的相关性分析，对产业布局管理和投资组合设计具有重要参考意义。韦颖璐 [88] 基于 CE 概念，提出了 pair-copula 熵的概念，用于度量多变量内部的成对相关关系。她将该概念应用于国内畜禽养殖产业链各环节之间的相关性研究，基于该领域内 9 家上中下游主要上市企业的股票价格数据，运用 pair-copula 熵度量了产业链内上中下游之间的相关性，发现该产业链上游相关性较强，下游相关性较弱；无条件相关性强，条件相关性弱；上中之间相关性强等现象。

投资者情绪对财经市场有着广泛而多面的影响，投资者情绪分析是经济学研究的重要问题之一。由于社交媒体和市场关系整合，投资者情绪会在人群和国家间传播，进而形成传播网络，使得局部情绪波动得以迅速扩散，造成系统性影响。Han 和 Zhou [89] 提出了一个基于小波分析、传递熵和网络分析组合的方法，研究公司间投资者情绪传播的模式，其中采用了基于 CE 的传递熵估计方

ChinaXiv:202105.00070v16

法。他们采用 2015-2021 年间的中国 137 家新能源汽车上市公司的百度搜索索引数据来代表投资者情绪，将其用小波分析分解为多尺度信息，再用传递熵构建情绪传播网络，最后用网络分析的方法分析短期和长期传播特征。他们发现，投资者情绪表现为短期局部活跃，并具有连续且逐渐增长的进化模式。

5.23 管理学

准确预测农产品期货价格有助于为政府相关部门的科学决策提供参考，因而对保障国家粮食安全具有重要意义。然而价格预测受多种复杂因素的影响，如国际形势、市场情绪博弈等。因此，识别价格的影响因素对构建准确的价格预测模型至关重要。An 等 [90] 提出了一个基于历史数据和文本数据的融合多种方法的混合预测框架，其中经验模态分解（Empirical Mode Decomposition：EMD）用于预处理历史数据，动态主题模型（Dynamic Topic Model：DTM）和情感分析用于提取微博文本信息，再利用 CE 等方法对提取的因子进行筛选，用于构建预测模型。作者在两个实际数据上验证了该方法框架：国家统计局的猪肉价格数据和大连商品交易所的大豆期货价格数据，并收集了相应时间内的微博文本数据。在实验中，作者将 CE 方法与同类的 dCor 和 HSIC 方法进行了对比，结果表明，在两个数据上，基于 CE 的预测模型都给出了最好的预测性能。

库存管理是企业运营管理过程中的关键环节，也是管理学的重要问题之一。报童问题是典型的单周期库存管理模型，一直是本领域研究的焦点。近年来，利用数据驱动模型和方法的报童问题研究展现出比传统方法的优越性，进而成为了热门话题。Tian 和 Zhang [91] 提出了一种端到端的算法框架，利用深度学习模型从在线商品评论等特征数据中预测订单数量，其中采用了包括 CE 在内的方法来选择模型的输入特征。他们将方法应用于汽车库存管理问题，基于 2016 至 2022 年间的大众朗逸汽车的历史销售量、某网站的评论、某搜索引擎指数、和宏观经济指数等数据构建了模型。结果显示，本方法能够大幅减少超额成本和短缺成本之和，与同类方法相比减少了 31.8% 的成本。

5.24 社会学

性别不平等是社会学研究的问题之一。由性别视角，我们可以发现很多不平等现象，如两性在收入上、教育上、职业上的不平等等。分析和鉴别导致不平等现象的社会学因素是学者们关心的问题，利用定量方法分析相关社会学数据是研究的手段之一。然而各种社会因素之间的因果链条十分复杂，需要采用科学的数据分析工具加以应对。马健 [11] 提出了一种多域因果关系鉴别方法，将性别因素作为社会外在变量，将不平等问题转化为数据分析中的域迁移问题，利用基于 CE 的条件独立性测试发现社会变量之间的因果关系。他将方法应用于美国国家成人收入社会调查数据，分析了性别、教育和收入之间的因果关系链条，发现了性别导致教育不平等，进而造成收入不平等的科学证据。

ChinaXiv:202105.00070v16

5.25 教育学

高中教育各学科之间具有内在的联系，教学大纲中强调了数学对物理、化学和生物等学科的基础性地位，数学知识、数学思维和思想方法深刻地渗透影响着其他学科的教学。因此，数学成绩被认为与其他学科成绩具有相关性。利用实证的方法研究数学与其他学科的关系，分析数学成绩与其他成绩之间的相关性是一个重要的基本问题，对于教学改革和学习方式的选择具有普遍参考意义。柳琼 [92] 基于某市 2013 级理科学生高一、高二期末考试成绩和高三两次模拟考试成绩，研究了数学成绩与其他学科成绩之间的相关性。作者比较了经典线性相关系数、秩相关系数和 MI 三种相关性度量方法，从 CE 和 MI 理论关系的角度分析论证了 MI 度量的优越性，并实验证明了 MI 度量能够更好地刻画揭示数学对其他不同学科（语文、英语、物理、化学和生物等）的影响力机制。

5.26 计算语言学

城市服务热线是政府公共管理系统的重要组成部分，促进了政府和市民的沟通，改善了政府的公共服务。但传统的人工派单方式无法满足日益增长的热线诉求，如何高效快速的处理大量的市民热线诉求是城市服务热线提高服务质量面临的重要课题。大量的热线文本数据积累为快速筛选和处理热线诉求提供了可能，可以利用自然语言处理方法处理热线文本数据，进而构建智能派单系统。陈作海等 [93] 提出了一种基于知识图谱技术的城市热线派单方法，基于城市热线数据构建热线知识图谱，再对待派单诉求根据构建的知识图谱检索结果进行派单，大大改善了热线服务的工作效率。在此智能派单系统中，CE 作为特征选择方法被用来对城市热线数据进行预处理，以构建和更新知识图谱。结果表明，CE 表现优于其他同类方法。作者将该方法应用在济南市民服务热线的系统上，通过不断更新知识图谱，最终获得了 90% 以上的派单准确率。

5.27 新闻传播学

公共卫生事件发生过程如何影响公众情绪是一个重要的问题，具有理论和现实意义，对政府的信息发布和舆情管控具有参考价值。特别是新媒体环境中，公众情绪的传播和演化过程受多种因素影响，因而更趋复杂。新冠疫情的发生给研究这类问题提供了条件。Zhang 等 [94] 研究了上海新冠疫情发生期间，疫情过程对公众情绪的影响特点和机理。他们以微博平台上“上海疫情”主题的数据为基础，研究了公众情绪的影响因素、时间演化以及疫情与公众情绪之间的因果关系。研究利用了基于 CE 的传递熵方法分析了疫情和公众情绪之间的因果关系，实证地发现了疫情过程对公众负面情绪的因果效应大于正面情绪，且正面情绪对负面情绪具有抑制效应。

ChinaXiv:202105.00070v16

5.28 法学

社区是基本的社会生活单元，社区治安管理与每个人的生活息息相关。社区属性与社区犯罪之间具有内在联系，分析社区经济、社会和人口等属性与各类犯罪之间的关系，可以加深对犯罪行为发生的理解，对执法部门合理安排部署资源力量具有重要参考意义。Wieser [95] 基于 CE 与 MI 的等价关系，提出了一种新的信息瓶颈（Information Bottleneck）估计方法。由于利用了 CE 的变换不变性，该方法较传统同类方法具有更好的估计性能。他将该方法应用于美国社区与犯罪数据集，分析 125 种经济社会因素与 18 种犯罪属性（包括 8 种犯罪行为，人均犯罪率和人均（非）暴力犯罪率）之间的关系，学习得到了可以表示这种关系的潜变量模型，为构建犯罪预测模型提供了参考。

5.29 政治学

政治安全事关国家安危。政治学研究关心政权领导力因素与政权危机之间的关系，并根据这些信息配置资源，开展情报收集、稳定或颠覆政权等行动。基于雪城大学莫伊尼汉全球事务研究所的国际政治领导力数据集，Card [96] 研究了 37 个领导力因素与政治安全之间的非线性关系，采用 CE（MI）作为非线性分析工具，重点关注了两个领导力变量（政权建立原因和政权结束原因）与其他因素的关系。分析结果佐证了社会学家的已有理论，分析也印证了已知的关系，发现了未知的关系和现象。

5.30 军事学

目标意图及时准确识别是战场态势感知的一项重要内容，是指挥决策的基础和前提。空中飞行目标意图识别会面临多种不确定性的挑战，如行为特性与物理特性的不确定性、飞行规则的不确定性和行动能力的不确定性等，使得及时准确的意图识别十分困难。张可等 [97] 提出了一种基于动态贝叶斯网络的目标意图识别方法，用于从复杂态势中目标的时序数据中完成意图识别，方法利用基于 CE 的 MI 估计算法从目标属性和目标意图数据来生成贝叶斯网络结构，再利用自适应遗传算法迭代优化网络结构，利用最终优化得到的网络来进行未知目标的意图识别。他们将该方法应用于空中目标的处理过程，利用空中目标的位置信息、飞行信息，以及雷达和通讯系统信息来识别其 6 种不同意图（巡逻、预警/指挥、电子侦察、电子干扰、攻击和打击等）。该方法可不限于空中飞行目标，可以很方便地推广到其他类型目标上。

5.31 情报学

颠覆性技术是具有原始创新性的技术，会对现有主流技术和产业产生变革性作用，推动经济社会发生突变式进步。开展颠覆式技术的前瞻识别及预判研

ChinaXiv:202105.00070v16

究是科技情报分析领域的重要问题，对科技政策制订、科技产业布局和科技创新生态培育具有指导意义。基于知识网络分析的科学、技术和产业互动模式研究是解决识别研判问题的路径之一。许海云等 [98] 提出了一个颠覆性技术研究流程框架，以渐进式技术为参照获取科技、专利和产业文献资料的文本数据，利用自然语言处理技术分别构建三者的知识网络，再利用知识网络的三种整体网络属性和网络社区相似度属性将知识网络互动模式划分为预设的五种模式，包括科学-技术-产业联动模式。其中，CE 被用来度量三种知识网络的整体网络属性之间的关联度，以表征互动模式。他们以再生医学（干细胞）领域作为颠覆性技术对象，以白血病治疗领域为渐进性技术参照开展实证研究，获取了截至 2020 年底的权威数据库相关文本数据，利用该流程框架研究了两个对比领域科学-技术-产业互动模式的共性和差异，加深了对颠覆性技术创新生态要素的知识流动和扩散规律的认识。

5.32 能源工程

天气是能源系统的重要影响因素，直接影响能源的生产和消费两端。特别是当可再生能源整合到能源系统中后，风速和光照等天气因素决定了风能和光伏能源的生产能力，而温度变化则会影响居民的能源消耗需求。但自然系统具有较大的随机性，给新能源系统的稳定高效运行带来了挑战。因此，新型能源网络管理系统需要建立合理的模型，以便将新能源集成到网络中。信息论为管理天气系统的随机性提供了工具。Fu 等 [99] 研究了基于信息论在集成能源系统中建立天气模型的方法。作者采用了 Copula 函数建立天气变量的联合分布模型，并采用 CE 计算的 MI 作为模型准确性的评价指标，以指导建模过程。同时，MI 还被用来衡量各种能源产出之间的关联强度。作者将得到的集成能源系统模型用于模拟中国北方某地区的能源系统运行情况，并与实际数据进行了对比。结果显示，系统模型的模拟与实际情况基本符合，说明构建的天气模型能够满足能源管理系统运行需求。

光伏发电技术受天气等环境因素影响，具有较大的不确定性，给电网的安全稳定运行构成影响。根据气象条件等因素对光伏发电站有功功率进行预报，有助于电网调度人员更好地制定调度策略，应对光伏发电的不确定性给电网的冲击威胁。朱正林和张冕 [100] 提出了一种结合优化算法、模态分解、CE 和深度学习模型的方法，用于提高发电功率的预测精度。他们在澳大利亚 Yulara 地区光伏电站数据上将方法与多种同类方法进行了对比，表明该方法得到的模型能够更好地适应天气变化的影响，取得最好的预测效果。

风能作为一种主要的清洁能源，具有间歇性和不确定性的特点，导致风电机组的功率预测和控制十分复杂。基于风电机组的监测数据，分析机组内各变量之间的相关性特征，有助于机组的健康状态监测和风电功率预测，从而更好地利用风能资源。崔双双和孙单勋 [101] 提出利用 CE 来分析风电机组状态变量之间的相关性，再基于 CE 相关性进行聚类以得到机组工况的划分。他们将方

ChinaXiv:202105.00070v16

法应用于广东某海上风电场数据采集与监控（SCADA）系统的数据，发现 CE 方法较传统方法能更好地描述数据中的相关性，并利用 K-means 方法得到了能精确地反映风电机组运行特性和状态的工况划分，具有重要的现实意义。

电力负荷预测是根据历史数据来预报未来一段时间的用电量，对智能电网调度和规划电力输送具有重要意义。电力负荷受多种因素影响，具有周期性和季节性等特点，特别是受天气因素的影响明显。因此，构建准确的电力负荷预测模型需要考虑天气等多种因素，并对天气对负荷的影响特点进行分析。Ma [14] 提出利用基于 CE 的 TE 方法来分析动态系统的时延特性，并将方法应用于摩洛哥缔头万 (Tétouan) 城的电力消费数据，从时延的角度分析了五种天气因素对该城三个电力供应网络的负荷的影响，发现了影响的每日时延变化特征。

可再生的风光能源越来越成为电力能源的重要组成部分，如何保证风光电力接入的经济效益和安全可靠是可再生能源利用的主要关切。合理的规划对于解决此关切十分关键，可保证建设投资回报和系统合理运行，防止风光能源被弃用的发生。储能系统可以平抑风光能源的不稳定波动性，是风光系统规划的组成部分。董海燕等 [102] 提出了一种考虑源荷时序相似性的风光储协同规划配置方法，其中利用 CE 衡量风光能源与负荷之间的相似性，以提高系统风光能源的利用效率。他们将方法应用于某工业园区的风光火储联合发电系统的规划配置，结果表明，该方法能有效降低储能系统的装机容量，提高新能源的消纳能力，经济效益和减排效益明显。

频率是电力系统最重要的物理量指标之一，频率稳定性是保障电力供应稳定性的一个基本要求。可再生能源由于具有不可预测性，其大量接入电网给电网频率稳定性带来了挑战。为了稳定和控制新能源带来的频率波动，需要准确快速地预测系统的频率稳定性，以帮助系统操作员提前制定控制策略。传统的频率稳定性预测是模型驱动的，由于求解耗时从而无法做到在线预测。基于机器学习的模型方法，通过简化模型以提高计算效率，可以满足在线预测的需求。Liu 等 [103] 提出了一种结合深度学习和 CE 的频率稳定性预测方法，CE 被用来选择模型输入变量，减少冗余信息以提高计算效率。作者将方法应用于两个系统：一个是新英格兰 39 节点系统，集成了美国西部电力调度委员会的动态风场模型；另一个是基于南加州西部的电网系统建立的 ACTIVSg500 系统。实验表明该方法建立的模型相较同类模型取得了最好成绩，达到了实用的要求。CE 方法不仅简化了模型、大幅降低了计算时间，且分析发现了与频率稳定性相关的电网变量，使得模型具有了可解释性。

电力系统宽频振荡由变流器控制主导，激发机理复杂，具有显著的时变、非线性和广域传播等特征，难以有效地进行建模分析。冯双等 [104] 利用 CE 的模型无关特性，提出了一种宽频振荡影响因素和传播路径分析方法。该方法以系统运行的状态参数为随机变量，通过计算其与各个频率区间的振荡阻尼之间的 CE 来选取影响振荡的关键因素；同时，利用系统发生振荡时的数据，计算系统变量之间的 copula 传递熵网络，用于分析振荡的传播过程。该分析方法是数据

ChinaXiv:202105.00070v16

驱动的方法，可以在系统模型未知的情况下得到相应的分析结果。

线损率是电力能源企业的一项综合性的经济技术指标，决定着其经济效益水平的高低。因此，线损管理和异常线损稽查是电力部门的一项重要工作。线损分析是利用科学的计算手段分析线损在电网中的分布规律，能为管理提供高效、准确的决策支持。Hu 等 [105] 提出了一种基于 TE 的线损分析方法，通过 CE 估计计算每个用户对区域总线损的 TE 值来判断其对总线损的贡献。他们基于每日电力供应和线损数据的计算分析，将用户根据线损贡献度排序，以应用于实际线损管理工作中，从而减少总线损率。

电价预测问题在电力市场参与者决策中至关重要，可以帮助其开发交易策略并合理分配资源。但新能源的广泛使用使电力供应具有不确定性，从而使电价预测变得更加复杂，造成预测模型构建较为困难。Xiong 和 Qing [106] 提出了一种基于时序数据的混合电价预测框架，将基于 CE 的特征选择方法与信号分解、贝叶斯优化和 LSTM 模型相结合, 以构建预测模型。他们将方法应用于2017 年美国宾夕法尼亚州-新泽西州-马里兰州互联网络（PJM）电力市场数据上，证明了该方法的有效性和实用性。

锂电池是使用最广泛的绿色清洁能源。但锂电池的电池容量会随着使用次数而退化，因此电池健康状态监测是电池管理系统中的主要问题之一。传统的健康状态监测模型大多在单一负载状况假设下得到，无法适用于真实场景下的多种状况，导致在原始数据上得到的模型无法适应新的情况。针对此问题，Hu 和 Wu [107] 提出了一种基于迁移学习思想的电池容量估计方法，结合了因果分析、注意力机制和 LSTM 等工具，其中基于 CE 的 TE 被用于选择与容量退化相关的健康状态指标，以保证构建模型在不同状况下的可迁移性。作者将方法应用于 NASA 的 3 种负载状况下的锂电池退化数据，结果表明，基于因果分析构建的模型比基于两种传统方法的模型的跨工况预测准确度分别提高了 8.6% 和 12.4%，增强了模型的鲁棒性。

5.33 食品工程

葡萄酒作为一种奢侈农产品，越来越走进广大普通消费者。葡萄酒质量的品鉴对其生产和销售都至关重要，葡萄酒酿造业大量投入在质量评价环节，以改善酿造工艺并促进消费。传统的质量品鉴主要依靠理化测试和专家感受，但专家的味觉感受主观性较强，其内在机理难以理解。因此，有必要研究酒的成分和专家评价之间的内在联系，以增进对葡萄酒质量的理解，提高质量评价的客观性。Lasserre 等 [108,109] 利用基于 CE 的（条件）独立性度量估计，提出了一种因果关系网络学习算法，称为 CMIIC，并将其应用于著名的葡萄牙绿酒的质量评价数据上，分析发现了分别与红葡萄酒和白葡萄酒的质量相关的理化成分。

ChinaXiv:202105.00070v16

5.34 土木建筑

建筑能源消耗占全部能源消耗的四成左右，建筑节能技术是重要的绿色能源技术，对实现联合国的碳中和目标意义重大。供暖、通风和空调（HVAC）系统贡献了商业楼宇四成以上的能耗，是建筑节能的主要研究对象之一。HVAC 系统的运行具有时延的特性，来自于媒介传导的滞后和热惯性。理解并运用这种特性，有利于设计适当的控制策略，从而达到节能的目的。Li 等 [110] 将基于 CE 的 TE 理念方法引入到 HVAC 领域，开发了一种基于信息论框架的无模型时延鉴别方法，用于 HVAC 系统的时序预测。他们改进了 kNN 的多变量 TE 估计器，结合优化方法设计了时延鉴别算法。他们将算法应用于大连某四层教学楼的供热监控系统，分析室内温度与天气参数（如室外温度、相对湿度、太阳辐射、风速等）和供热参数（如热水供应和回流温度等）的数据，鉴别时延特性，进而利用后两组参数预测下一段时间的室温。结果表明，TE 方法能够鉴别参数之间的时延关系特性，进而提高室温预测性能。

工程变形监测是工程测量领域的重要问题之一，需要保证监测精度和可靠性，对大型工程的施工运营安全具有重要意义。常见的变形监测分析方法一般只针对单个监测点的建模和预测，但变形体内部监测点间不是孤立的，而是具有内在的相关性，因而可以利用这种相关性提高单点监测的预测精度。曹久慧等 [111] 提出了一种基于自注意力机制的变形监测方法，采用 CE 度量监测点与周围点之间的相关性，再利用这些相关点组成的数据集训练自注意力机制的模型来进行长时间的变形预测。他们将方法应用于某隧道施工段 2020 年 12 月至 2021 年 10 月间围堰上监测点位的位移数据，以预测 7 天的变形，获得了较好的效果，对实际工程围堰预警等长期变形预测问题具有良好的应用价值。

5.35 交通运输

大件货物运输是指通过多种运输方式对具有不可拆解属性的大型物件的专业运输作业活动，在国民经济中占有重要地位，对国计民生重点行业的基础设施建设起着重要的支撑和保障作用，也关系着国防军事和国家安全。大件货物运输大都需要铁路、航运等多式联运的方式才能完成，需要制定各个局部运输环节模块联动的整体方案。随着交通系统的数字化，大量的相关方案数据得到积累，基于数据的大件货物运输方案制定成为了一个重要的问题，其研究有助于提高方案制定的科学性和适用性。黄达 [112] 利用 CE 等多种数学工具提出了一种基于模块链构建的大件货物多式联运方案制定方法。该方法先将运输方案分解为多个局部环节模块，再利用 CE 等相关性度量工具筛选一组模块属性用于计算方案之间的相似度，最后在已有运输案例库中检索与目标运输任务相似度高的案例作为初步运输方案。由于大件运输方案的多样性，一些案例模块属性会具有非高斯性，使得传统的相关系数工具不再适用于计算属性间相关性，而 CE 由于具有普适性则依然适用。作者在 600 多个实际案例的数据上验证了

ChinaXiv:202105.00070v16

该方法，并构建了方案制定原型系统。

航空和高速铁路是我国最主要的两种旅客运输方式。相较于航空，高铁票价的市场化水平处于落后的水平，欠缺灵活性和动态性。因此，研究影响票价的因素以期改进高铁票价的定价机制是学界十分关心的问题。许罗豪等 [113] 基于京沪航空和高铁票价的数据，利用 CE 和决策树等工具研究了出行需求、旅客选择、出行效率和出行路线四类因素对航空和高铁票价的影响。他们发现购票提前期对两种票价的影响程度不同，但旅行时间对二者的影响程度较为相似。这些研究结论对高铁定价具有一定的参考价值。

5.36 制造工程

产品质量是制造业的生命。注射成型（injection molding）是近年快速发展的工业制造技术，在航天、建筑、通讯等领域有着广泛应用。注射成型过程包括了多步复杂的物理和化学反应过程，很容易受到外部因素的影响，保证塑料产品质量的稳定性是一个难题。基于制造过程历史数据，建立产品质量预测模型是提高产品质量的手段之一。但建立模型需要首先选择有关的过程参数作为模型输入，以获得较好的预测性能。Sun 等 [114] 提出基于 CE 方法选择过程参数变量用于构建质量预测模型，并将方法应用于真实的富士康公司的注射成型生产过程数据，大幅改善了质量预测的性能。

复杂机械产品的整机制造包括设计、制造和装配三个环节。作为产品生产的最后一个环节，装配过程在零部件的制造过程基础上组装高精度产品，装配质量控制在零部件制造质量的基础上保障整机产品质量。复杂机械产品零部件数量种类繁多、相互关联，装配环节错综复杂，上游环节的装配质量误差会对下游环节质量构成影响。王小巧 [115] 在装配质量控制中考虑了上下游工序和质量控制点之间的相关性，利用 Copula 对控制点间相关关系建模，并用 CE 度量这种相关性，进而提出了一种装配质量控制点控制阀优化方法。她将方法应用于江淮汽车某型汽油发动机关键零部件缸盖的装配工序过程，验证了方法的有效性。

现代工业系统变得越来越高度复杂和自动化，使得工业过程监测变得愈加困难。如何监测系统异常并发现异常原因是一个具有广泛应用的重要问题。利用因果分析得到工业系统内部复杂的因果关系图，有助于准确发现异常的传播路径，进而及时进行干预。Dong 等 [116] 提出了一个结合动态 PCA、TE 和 LSTM 的故障分析框架，其中基于 CE 的 TE 被用分析系统内的因果关系。作者将该方法应用于辽宁鞍钢的热轧带钢工艺过程数据的分析，成功地对过程中的两个故障及其原因进行了分析。作者还将基于 TE 的因果图分析方法与同类格兰杰因果分析方法进行了对比，表明 TE 方法能够更准确地对故障进行根因分析。

烧结过程（Sintering Process：SP）在钢铁工业中至关重要，同时也会消耗大量的能源。动态预测 SP 的碳消耗有助于节约能源和减少碳排放。传统的 SP

ChinaXiv:202105.00070v16

建模基于一定的假设，无法适应 SP 的系统动态特性，基于数据的机器学习模型可以克服传统模型的不足。Hu 等 [117] 提出了一种动态建模方法框架，可以自动识别过程工况状态，从而进行碳消耗预测。该方法框架结合了 AKFCM 聚类算法、基于 CE 的模型选择和宽度学习模型方法。作者在一家钢铁企业的实际数据上验证了方法的有效性，证明了 CE 可以快速地捕捉不同工况下 SP 中复杂的相关关系模式，从而使该方法能够比传统方法更准确地预测烧结碳消耗。

5.37 可靠性工程

退化过程（degradation processes）在各种工程系统中普遍存在，导致系统可靠性的降低甚至失效，如金属材料的疲劳和腐蚀、半导体器件的参数漂移等。退化过程建模是评估系统和产品有效性和寿命的主要技术手段之一。由于现代系统的复杂性，影响退化过程的因素较多，因素变量本身具有非线性特征，且变量之间又相互关联，从而对退化过程建模构成了可靠性工程的一个基本难题。如果建模时忽略了因素之间的相关性，就会导致模型错误和可靠性估计误差。传统的衡量因素之间的相关性主要采用线性相关系数，难以处理复杂的相关关系。Sun 等 [118] 提出采用 copula 对过程因素之间关系建模，并用 CE 来度量退化过程因素之间的关联。他给出了一种参数化 CE 估计方法，并成功应用于微波电子组件的退化过程分析中。结果表明，该方法能够分析不同阶段的退化过程。

5.38 化学工程

故障诊断对化学过程的安全、高效运行至关重要，数据驱动的故障诊断方法是实际生产运行中的主要方法之一。为了构建诊断模型，构建合理的正常和故障状态的过程表示是问题的关键环节。Yin 等 [119] 提出了一种基于 CE 的灰度相关空间的故障诊断方法，通过变量之间的 CE 相关性矩阵来刻画过程的正常和故障状态，再将矩阵作为卷积神经网络的输入来构建故障分类模型。他们将方法应用于田纳西伊斯曼（Tennessee Eastman）过程的故障诊断数据，结果表明该方法取得了 95% 以上的诊断准确率，验证了方法的有效性。主元分析法（PCA）是一种常用的多变量过程检测方法，原理是基于最大方差准则从一组过程变量构建过程检测统计量，但其仅适用于线性的情况。Wei 和 Wang [120] 提出了一种基于 CE 的非线性 PCA 方法（CEPCA），从具有非线性特征的 CE 矩阵得到过程检测统计量。他们将方法应用于田纳西伊斯曼过程数据，并与 PCA 方法进行了对比，结果表明，CEPCA 方法获得了更好的故障检测率结果。

理解化工过程变量之间的因果关系对于过程控制十分重要，有助于更好的过程监测和故障诊断。利用因果发现方法构建化工过程因果关系图，可以对故障进行根因分析，是故障诊断的重要方法之一。Bi 等 [121] 提出了一种基于深度学习进行因果发现的 CGTST 方法，并与基于 CE 的 TE 等多种方法进行了对比。实验结果表明，在一个 5 变量的连续搅拌槽式反应器数据上，TE 方法获

ChinaXiv:202105.00070v16

得的反应图结果非常接近于真实情况；在田纳西伊斯曼过程数据上，TE 方法也取得了接近于真实情况的估计结果，体现出了较强的实用性。

5.39 航空航天

航空飞行器系统日趋复杂，飞行器设计首先需要加深对其总体设计参数的认识。对各种设计参数间的耦合关系的理论分析，有助于分析设计方案可行性或优化总体设计方案。Krishnankutty 等 [122] 基于 CE 与 MI 的等价关系，提出了两种基于 Copula 的 MI 估计方法，并将方法应用于美国 22 种喷气战斗机的技术参数数据的分析，估计了飞行航程和可承受负载之间的耦合关系，验证了分析方法的有效性。

卫星是航天时代的主要航天器类型，在信息时代有着广泛的民事和军事用途。作为一种在极端环境运行的复杂系统，卫星的在轨健康状态监测十分重要。卫星遥测数据是各种传感器参数的编码，包含了卫星内部运行系统物理参数的交互关系信息。卫星的异常模式会由于这种交互而在内部传播，因此分析这种内部交互导致的故障传播链条有助于及时发现卫星异常状态，保障卫星正常运行。分析遥测参数之间的因果关系是一种解决问题的路径。Liu 等 [123] 提出直接将基于 CE 的 TE 应用于分析真实的卫星遥测数据，得到了遥测参数之间的故障传导图，结果要优于传统的 TE 方法。Zeng 等 [124] 提出了一种改进的 TE 度量，称为 NMCTE，用于分析遥测参数之间的因果关系网络，该度量利用了基于 CE 的 TE 表示和估计方法。他们又提出了基于所得因果网络的异常检测的 CN-FA-LSTM 方法。他们将 NMCTE 方法应用于真实的卫星遥测数据，得到了具有良好的可解释性的因果网络。他们又将 CN-FA-LSTM 方法在 NASA 公开的 SMAP 和 MSL 数据集上与其它 6 种方法进行了对比，验证了方法的优越性。

航班延误是影响国际民航业正常有效运行的主要问题之一，不仅给旅客造成出行不便，也给航空业带来巨大经济损失。航空系统是一个有机的整体，运行中存在航班资源的上下游共享，带来了系统耦合，导致上游航班的到港延误会向下游传播，因此航班延误管控首先需要对这种延误因果关系进行分析。吴格等 [125] 提出利用一种基于 CE 的 TE 估计器来分析机场的航班延误时间序列之间的因果关系强度的方法，使民航信息系统具有了分析两个航班之间是否具有延误因果关系的能力，从而能够深入理解和利用航空系统节点间航班延误的内在关系。

5.40 车辆工程

现代汽车的电子设备系统由车载网络连接集成，提高了乘坐的舒适性、安全性和多功能特性。但随着智能车辆技术的发展，车内设备也成为了黑客攻击的对象，对车辆安全构成了威胁。CAN 总线是一种智能车辆内连接控制各个车

ChinaXiv:202105.00070v16

辆电子组件的数据通信协议，已在汽车领域成为事实上的主流标准，但由于缺乏加密、认证等机制，其在网络攻击面前非常脆弱。因此，研究 CAN 总线的入侵检测技术成为了提高其安全性的主要技术手段之一。Gao 等 [126] 提出了一种轻量级神经网络设计方法，用于检测 CAN 总线入侵事件，其首先分析异常 CAN 数据帧的属性集合，再利用 CE 选择出众多属性中与入侵攻击有关的少数属性，再利用这些属性构建一种 CanNet 神经网络检测器以检测入侵。他们利用现代汽车索纳塔 YF 的 CAN 总线数据验证了 CanNet 方法，结果表明该方法与同类方法相比具有高检测率、高实时性和低内存占用的优点。

5.41 电子工程

半导体芯片的集成度的不断提高，对微电子封装的要求也越来越高。微电子封装起着隔绝外部环境、散发内部热量的功能，对集成电路的稳定运行具有至关重要的保护作用。这就要求封装材料具有良好稳定性、高强度，同时还要满足其他物理性质。刘勃 [127] 以 Cu 基材料为主体，建立 CuNi 二元合金体系，利用第一性原理与机器学习相结合的方法，基于团簇相关函数特征，预测分别与材料强度和稳定性相关的构型能和杨氏模量。作者利用 CE 分析了预测模型的合理性，通过计算特征之间的相关性，以及特征与构型能和杨氏模量之间的相关性，发现模型特征与杨氏模量之间的相关性更高，同时构型能与杨氏模量之间的相关度较低，增进了模型的可解释性，有助于设计更合理的材料性质预测模型。

5.42 通信工程

通信安全是移动通讯的主要关切之一，一般通过通信层的加密技术加以解决。在资源受限的新兴网络（如 IoT、WSN 等）中，密钥分发是一个挑战。无线信道的互易性为通信双方提供了共享密钥的机制，双方可通过测量无线信道获取密钥。密钥容量概念为无线信道密钥提取提供了理论上限。然而，现实中密钥容量往往受到诸多实际物理条件（如终端移动、信道噪声等）的限制，需要对其进行定量分析。Wang 等 [128] 研究了均匀散射环境下物理因素对密钥容量的影响，将其转化为随机变量的 MI 计算问题，并基于仿真物理环境验证其理论推导的正确性，仿真实验采用了基于 CE 的 MI 估计算法估计密钥容量。仿真结果表明，理论推导得到了验证，能够指导实际应用。

第 6 代（6G）通信网络技术的研发需要面对的主要挑战之一就是要达到更高的数据传输率，以满足更极致的体验、3D 视觉、工业智能等场景需求。传统的通信理论没有考虑传输信息中的语义信息，而 6G 技术可以利用基于 AI 的语义通信来达到更高的网络传输性能。傅宇舟等 [129] 提出了一种面向 6G 网络的基于语义通信的端到端服务框架，将语义通信与 AI 的语义分析能力相融合，利用基于 Transformer 的编解码器来压缩语义信息。其中，语义编码器的损失

ChinaXiv:202105.00070v16

函数由基于欧式距离的语义损失函数和基于 CE 的信息量损失函数组成。他们利用图像数据验证了该服务框架，使用 ImageNet-1K 数据集训练框架，再使用 VOC2012 数据集进行仿真验证。结果表明，与传统通信方案相比，该服务框架在目标检测和图像语义重建上均取得了最优性能，且取得了与全语义特征传输方案相近的性能，有望成为 6G 网络的技术内容。

5.43 高性能计算

提高能源效率是高性能计算研究的一个重要目标。通过配置程序的最优能效设置，如处理器频率等，可以降低程序执行时的能耗。但决定最优配置是一个费时的过程，程序一旦修改就需要重新配置。利用机器学习方法通过性能事件来自动决定最优配置是一个新的研究方向，但需要确定哪些事件是能效相关的以决定最优配置。Gocht-Zech [130] 提出利用特征选择的方法来选择能效相关事件，他选择了 6 种特征选择方法，并基于 CE 理论给出了相应的估计方法。实际数据实验表明该基于 copula 的方法能够鉴别出能效相关的性能事件，从而提高程序执行时的能效，在增加 7% 运行时的成本下节省了 24% 的能源消耗。

5.44 测绘遥感

高光谱遥感是应用广泛的前沿测绘技术，通过遥感光谱成像，能够获取不同地物的诊断性光谱信息。由于高光谱图像波段数多，数据大且存在大量冗余信息，需要利用特征提取技术对有效波段进行选择，以表征成像对象体。因此，高光谱图像波段选择是该领域的重要问题之一，主要思想是选择一个波段子集，使得成像评价准则函数达到最大。其中，基于信息论的准则是波段选择的主要方法之一。Zeng 和 Durrani [131] 提出利用基于 CE 的 MI 选择波段的方法，并将其应用于美国印第安纳西北的 Indian Pine 处采集的真实高光谱数据，结果表明 CE 提供了一种鲁棒的 MI 波段选择方法。

5.45 金融工程

量化金融是通过对金融数据的数量关系分析指导金融决策的新兴金融学科。基于金融交易系统产生的大量金融市场交易数据，利用数学工具分析金融产品之间的数量关系，可以明晰市场规律和动态，进而管理金融资产。其中，分析市场金融变量之间的相关性是金融工程的重要问题，可以帮助交易员洞察它们之间的动态关系，进而调整投资组合和管理风险。由于金融市场变量具有非线性、非高斯性等特征，使得 MI 成为了理想的相关性度量，而 MI 估计算法则成了量化金融工具箱的重要工具之一。基于 CE 的 MI 估计算法就被量化金融算法库 `MLFinLab` [132] 实现，并得到业界广泛应用。

基于中国股票市场（沪市 A 股指数、深市 A 股指数和沪深 300 指数）真实数据，Wang [133] 研究了利用股票资产之间的相关性关系网络，优化投资组合

ChinaXiv:202105.00070v16

的方法。方法采用了包括 CE 在内的线性和非线性相关性度量，基于相关性强度构建股票资产间的关系网络，进而构建投资组合。研究中估计了不同 Copula 参数函数族的 CE（MI）。

股票市场的投资者总是希望投资发展良好的上市公司，因此甄别一只股票的好坏对投资者十分重要。ST 股票制度是在我国 A 股市场实施的股票风险警示机制，有助于投资者选择投资组合并规避风险。股票分类是股票分析领域的一类重要问题，对金融市场投资者具有参考价值。朱仲儿 [134] 提出了一种基于机器学习方法的 ST 股票分类方法，采用 Boruta 算法和 CE 方法进行特征选择，再利用 6 种回归模型进行预测，利用 Optuna 框架对模型的超参数寻优。他选取了 tushare 数据库中上交所和深交所的 2076 只股票（含 351 只 ST 股票）自 2016 年以来的数据，含有 139 个股票特征变量，最终利用 Boruta 和 CE 方法筛选了 7 个可解释的变量。模型预测结果表明该方法在筛选特征和 XGBoost 模型组合上获得了最好的预测精度。

分析金融数据需要对其建模数学模型，但金融变量以及其联合分布具有非高斯性，给数据建模带来了挑战。Calsaverini 和 Vicente [135, 136] 给出了一种巧妙的 Copula 函数模型选择方法。该方法利用 CE（MI）的边缘分布无关特性，将 Copula 鉴别问题的目标与边缘函数分开，再利用 CE 的定义，将问题转化为以 MI 为上界的模型选择问题。作者还定义了超量信息（Informaion Excess）的概念。作者将建模方法应用于 1990 至 2008 年间标普 500 指数的 150 只股票的每日对数收益率数据，利用超量信息，验证了该方法作用于 T-Copula 函数族时的有效性。

R 藤 Copula 是一种灵活的构建多元 copula 分布的工具，确定藤的结构是建立此类模型的关键步骤。Alanazi [137] 基于 CE 和 MI、CMI 之间的关系，提出了一种 R 藤 copula 的构建方法，基于 MI 建立最小生成树，再计算前一子树每对边上的 CMI，根据 CMI 建立新的子树并决定藤 copula 的层级结构。他将该 R 藤 copula 构建方法应用于股票间相关结构的建模问题，基于德国 DAX 指数 15 种主要股票数据（2005 年 1 月至 2009 年 8 月）构建了资产间关系结构的 R 藤 copula 模型，与传统方法相比，该方法建立的 copula 相关结构模型能够更好地拟合数据。

金融危机的发生使金融系统的系统性风险问题受到各国监管部门的关注。随着我国股票市场逐渐放开管制，经济金融一体化程度不断加深，各个行业间的耦合关系加大了系统性风险的程度，因此需要对跨行业的风险溢出效应加以研究，以期进行防范和化解。熵作为量化不确定性的数学工具，十分适合度量金融风险组合。熊靖宇 [138] 采用 CE 等工具对 2005 年 1 月 5 日至 2020 年 7 月 3 日我国股票市场 11 个行业的日对数收益率数据进行了分析，研究行业个体风险和跨行业风险溢出特征的动态演变过程，特别针对 2008 年金融危机、2013 年钱荒和 2015 年股灾三个时期的风险特征进行研究。研究发现，行业联合 CE 动态变化滞后于累加独立熵发生，说明了行业间联动导致了系统性风险增强；2008

ChinaXiv:202105.00070v16

年金融危机的市场内部传染性更强，破坏程度更大；近期 11 个行业内部关联水平较强。

金融脆弱性是由金融部门自身高负债经营带来的内在不稳定性。金融脆弱性度量工具可以使国家及时地对危机进行响应和干预，因此得到了大量的研究。日益成熟的网络分析理论为从金融网络的角度度量金融脆弱性提供了方法工具，但传统的网络构建方法只是基于线性关系度量工具，如皮尔逊相关系数等，不能够反映金融系统中的非线性关系特性。Chen 等 [139] 提出了一种利用 CE 改进的网络曲率（Network Curvature）金融脆弱性度量方法，该方法先利用 CE 构建金融网络，再计算网络的四种离散 Ricci 曲率作为市场脆弱性度量。他们将该度量方法应用于 2006 年 4 月至 2022 年 4 月间沪深 300 指数的股票数据，分析金融危机前后的市场脆弱性。结果表明，该度量方法比基于皮尔逊相关的方法更清晰地描述了金融危机后市场的脆弱性，且具有传统风险度量同样的风险度量能力。

信用风险是金融银行业面对的主要基本风险之一，保障金融安全需要有效地管理信用风险。信用评分卡模型是一种对客户进行信用风险评价的模型方法，是管控金融风险的决策工具。该类模型根据客户的信用历史数据为其划分信用等级，来决定其金融权限。传统的建立信用评分卡模型方法依靠专家经验，效率低且生成的模型不够完善。孔祥永等 [140] 提出一种基于 CE 的自动化信用风险模型构建方法，能够显著提高建模效率，可以同时保证模型具有高预测性能和可解释性。作者将该方法在真实信用卡数据上与专家建模进行了对比，实验结果表明方法大大缩短了建模时间，且能够得到媲美专家模型的预测性能和可解释的客户信用特征。

6 总结

统计独立性是统计学和机器学习领域的基础性概念，如何表示和度量统计独立性是该领域的基本问题。Copula 理论提供了统计相关性表示的理论工具，通过将随机变量的边缘函数与表示统计关联性的 Copula 函数相分离，得到了表示任何关联性的数学形式。而 CE 理论则给出了度量统计独立性的概念工具，度量了 Copula 函数表示中所有的信息量，也就是相关性的强度。CE 是一种具有诸多公理性属性的理想的统计度量工具。

本文综述了 CE 的理论和应用，介绍了 CE 基本概念定义、与 MI 等价性的定理和推论，以及 CE 的性质。介绍了 CE 的非参数估计方法。本文介绍了 CE 研究的最新进展，包括其在统计学九个基本问题（结构学习、关联发现、变量选择、因果发现、系统辨识、时延估计、域自适应、正态性检验和双样本检验等）上的理论应用，讨论了前四个理论应用之间的关系，探讨了这四个应用对应的深层次的相关性和因果性概念之间的联系，并将基于 CE 的（条件）独立性度量框架与基于核函数和距离的相关性度量框架进行了对比，指出了本理论框

ChinaXiv:202105.00070v16

架在多个方面的理论优越性，又通过仿真和实际数据实验评估验证了 CE 的实际优越性。

本文综述了 CE 在理论物理学、理论化学、化学信息学、材料学、水文学、气候学、气象学、环境学、生态学、动物形态学、农学、认知神经学、运动神经学、计算神经学、心理学、系统生物学、生物信息学、临床诊断学、老年医学、精神病学、公共卫生学、经济学、管理学、社会学、教育学、计算语言学、新闻传播学、法学、政治学、军事学、情报学，以及能源工程、食品工程、土木建筑、交通运输、制造工程、可靠性工程、化学工程、航空航天、车辆工程、电子工程、通信工程、高性能计算、测绘遥感和金融工程等多学科领域的实际应用。基于 CE 带来的理论和计算上的优势，在这些应用中 CE 被用来分析和度量各种类型数据中的统计关联性或因果性，通过选择变量来建立模型，以及作为评价指标评价模型，均取得了良好的应用效果。CE 作为一种基础性工具，不仅为这些应用提供了基本的理论和方法，也为应用中各种新方法论的派生提供了可能。

A 软件实现

本文所述的 CE 估计算法、TE 估计算法、正态性检验和双样本检验的统计量的估计算法已在 R 和 Python 语言的 **copent** 算法包中实现 [141]，分别在 CRAN 和 PyPI 上共享：

- CRAN `https://cran.r-project.org/package=copent`；
- PyPI `https://pypi.org/project/copent/`。

相关源码见作者的 GitHub：`https://github.com/majianthu/`。

另，第三方实现的 CE 估计算法包括：

- R 语言的 **Cylcop** 包 [51, 194]；
- Python 语言的 **MLFinLab** 包 [132]；
- Julia 语言的 **CopEnt.jl** 包 [195] 和 **CausalityTools.jl** 包 [196]；以及
- Matlab 和 Python 语言的 **gcmi** 包 [56, 197] 等。

参考文献

[1] Karl Pearson. Mathematical contributions to the theory of evolution.—on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of The Royal Society of London*, 60(1):489–498, 1896.

ChinaXiv:202105.00070v16

[2] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[3] Harry Joe. *Dependence modeling with copulas*. CRC press, 2014.

[4] Abe Sklar. Fonctions de repartition a n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

[5] Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1):51–54, 2011. See also arXiv preprint arXiv:0808.0845 (2008).

[6] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[7] Jian Ma and Zengqi Sun. Dependence structure estimation via copula. *arXiv preprint arXiv:0804.4451*, 2008.

[8] Jian Ma. Discovering association with copula entropy. *arXiv preprint arXiv:1907.12268*, 2019.

[9] Jian Ma. Variable selection with copula entropy. *Chinese Journal of Applied Probability and Statistics*, 37(4):405–420, 2021. See also arXiv preprint arXiv:1910.12389 (2019).

[10] Jian Ma. Estimating transfer entropy via copula entropy. *arXiv preprint arXiv:1910.04375*, 2019.

[11] Jian Ma. Causal domain adaptation with copula entropy based conditional independence test. *arXiv preprint arXiv:2202.13482*, 2022.

[12] Jian Ma. Multivariate normality test with copula entropy. *arXiv preprint arXiv:arXiv:2206.05956*, 2022.

[13] Jian Ma. Two-sample test with copula entropy. *arXiv preprint arXiv:2307.07247*, 2023.

[14] Jian Ma. Identifying Time Lag in Dynamical Systems with Copula Entropy based Transfer Entropy. *arXiv preprint arXiv:2301.06037*, 2023.

[15] Jian Ma. System identification with copula entropy. *arXiv preprint arXiv:2304.12922*, 2023.

[16] Jian Ma. On thermodynamic interpretation of copula entropy. *arXiv preprint arXiv:2111.14042*, 2021.

ChinaXiv:202105.00070v16

ChinaXiv:202105.00070v16

[17] Michel A. Cuendet, Harel Weinstein, and Michael V. LeVine. The allostery landscape: Quantifying thermodynamic couplings in biomolecular systems. *Journal of Chemical Theory and Computation*, 12(12):5758–5767, December 2016.

[18] Mario Wieser, Sonali Parbhoo, Aleksander Wieczorek, and Volker Roth. Inverse learning of symmetries. In *Advances in Neural Information Processing Systems*, volume 33, pages 18004–18015, 2020.

[19] 田杰. 基于机器学习的耐热型含能材料设计方法研究. 硕士学位论文, 西南科技大学, 2023.

[20] Lu Chen, Vijay P. Singh, and Shenglian Guo. Measure of correlation between river flows using the copula-entropy method. *Journal of Hydrologic Engineering*, 18(12):1591–1606, 2013.

[21] Lu Chen, Vijay P. Singh, Shenglian Guo, Jianzhong Zhou, and Lei Ye. Copula entropy coupled with artificial neural network for rainfall–runoff simulation. *Stochastic Environmental Research and Risk Assessment*, 28(7):1755–1767, 2014.

[22] Lu Chen and Vijay P. Singh. Flood forecasting and error simulation using copula entropy method. In Priyanka Sharma and Deepesh Machiwal, editors, *Advances in Streamflow Forecasting*, pages 331–368. Elsevier, 2021.

[23] Xiao Li, Liping Zhang, Sidong Zeng, Zhenyu Tang, Lina Liu, Qin Zhang, Zhengyang Tang, and Xiaojun Hua. Predicting monthly runoff of the upper Yangtze river based on multiple machine learning models. *Sustainability*, 14(18):11149, 2022.

[24] Ran Mo, Bin Xu, Ping-An Zhong, Yuanheng Dong, Han Wang, Hao Yue, Jian Zhu, Huili Wang, Guoqing Wang, and Jianyun Zhang. Long-term probabilistic streamflow forecast model with "inputs–structure–parameters" hierarchical optimization framework. *Journal of Hydrology*, page 129736, 2023.

[25] 温云亮, 李艳玲, 黄春艳, and 张泽中. 基于 Copula 熵理论的干旱驱动因子选择. *华北水利水电大学学报 (自然科学版)*, 40(4):51–56, 2019.

[26] C.Y. Huang and Y.P. Zhang. Prediction based on copula entropy and general regression neural network. *Applied Ecology and Environmental Research*, 17(6):14415–14424, 2019.

[27] 黄春艳. 黄河流域的干旱驱动及评估预测研究. 博士学位论文, 西安理工大学, 2021.

[28] Lingling Ni, Dong Wang, Jianfeng Wu, Yuankun Wang, Yuwei Tao, Jianyun Zhang, Jiufu Liu, and Fei Xie. Vine copula selection using mutual information for hydrological dependence modeling. *Environmental Research*, 186:109604, 2020.

[29] P. Kanthavel, C.K. Saxena, and R.K. Singh. Integrated Drought Index based on Vine Copula Modelling. *International Journal of Climatology*, 2022.

[30] M. Mohammadi, M. Emadi, and M. Amini. Bivariate dependency analysis using Jeffrey and Hellinger divergence measures based on copula density estimation by improved probit transformation. *Journal of Statistical Sciences*, 15(1):233–254, 2021.

[31] Pengcheng Xu, Dong Wang, Vijay P. Singh, Yuankun Wang, Jichun Wu, Lachun Wang, Xinqing Zou, Yuanfang Chen, Xi Chen, Jiufu Liu, Ying Zou, and Ruimin He. A two-phase copula entropy-based multiobjective optimization approach to hydrometeorological gauge network design. *Journal of Hydrology*, 555:228–241, 2017.

[32] 王栋, 徐鹏程, 王远坤, and 吴吉春. 一种基于 Copula 熵的水文站网优化模型的优化方法, 2019. CN106897530B.

[33] Heshu Li, Dong Wang, Vijay P. Singh, Yuankun Wang, Jianfeng Wu, Jichun Wu, Ruimin He, Ying Zou, Jiufu Liu, and Jianyun Zhang. Developing a dual entropy-transinformation criterion for hydrometric network optimization based on information theory and copulas. *Environmental Research*, 180:108813, 2020.

[34] Heshu Li, Dong Wang, Vijay P. Singh, Yuankun Wang, Jianfeng Wu, and Jichun Wu. Developing an entropy and copula-based approach for precipitation monitoring network expansion. *Journal of Hydrology*, 598:126366, 2021.

[35] 徐鹏程, 仇建春, 李帆, 刘赛艳, and 蒋新跃. 基于高维 Copula 熵和克里金的站网优化方法, 2022. CN114595556A.

[36] 徐鹏程, 李帆, 张昌盛, and 仇建春. 基于 C-Vine Copula 熵多目标优化模型的水文气象站网优化研究. 中国农村水利水电, 2:16–21, 2022.

ChinaXiv:202105.00070v16

ChinaXiv:202105.00070v16

[37] 杨惜岁. 多目标准则下流域水文站网的优化与评价. 硕士学位论文, 武汉理工大学, 2019.

[38] Lu Chen and Shenglian Guo. *Copulas and its application in hydrology and water resources*. Springer, 2019.

[39] Xu Wang and Yong-Ming Shen. A framework of dependence modeling and evaluation system for compound flood events. *Water Resources Research*, 59(8):e2023WR034718, 2023.

[40] Longxia Qian, Yong Zhao, Jianhong Yang, Hanlin Li, Hongrui Wang, and ChengZu Bai. A new estimation method for copula parameters for multivariate hydrological frequency analysis with small sample sizes. *Water Resources Management*, 36(4):1141–1157, March 2022.

[41] 刘磊, 高超, 王志刚, 王晓艳, 章四龙, and 陈娜. 基于非线性相关性和复杂网络的径流相似性分区. *水科学进展*, 33(3):442–451, 2022.

[42] Victor Costa Porto, Francisco de Assis de Souza Filho, Taís Maria Nunes Carvalho, Ticiana Marinho de Carvalho Studart, and Maria Manuela Portela. A GLM copula approach for multisite annual streamflow generation. *Journal of Hydrology*, 598:126226, 2021.

[43] 黄朝君, 贾建伟, 秦赫, and 王栋. 基于 Copula 熵-随机森林的中长期径流预报研究. *人民长江*, 52(11):81–85, 2021.

[44] Zengchao Hao and Vijay P. Singh. Integrating entropy and copula theories for hydrologic modeling and analysis. *Entropy*, 17(4):2253–2280, 2015.

[45] Francesca Condino. *La divergenza di Jensen-Shannon nell'algoritmo di clustering dinamico per dati descritti da distribuzioni multivariate*. PhD thesis, Università degli Studi di Napoli Federico II, 2009.

[46] Jujie Wang, Wenjie Xu, Yue Zhang, and Jian Dong. A novel air quality prediction and early warning system based on combined model of optimal feature extraction and intelligent optimization. *Chaos, Solitons & Fractals*, 158:112098, 2022.

[47] Xiaoxuan Wu, Chen Zhang, Jun Zhu, and Xin Zhang. Research on PM2.5 concentration prediction based on the CE-AGA-LSTM model. *Applied Sciences*, 12(14):7009, 2022.

[48] 陈燕璇, 刘合香, and 倪增华. 基于 Copula 熵因子选取的 PSO-ELM 台风灾情预测模型. *气象研究与应用*, 40(2):7–11, 2019.

[49] 吴京鹏. 基于图嵌入表示的节点无特征网络链路预测研究. 硕士学位论文, 西北师范大学, 2022.

[50] 金秀章, 乔鹏, and 史德金. 基于 VMD-Bayes-Lasso 算法带误差补偿的火电厂 NOx 浓度软测量. *华北电力大学学报 (自然科学版)*, 2023.

[51] Florian H. Hodel and John R. Fieberg. Cylcop: An R package for circular-linear copulae with angular symmetry. *bioRxiv*, page 2021.07.14.452253, 2021.

[52] Francisco Escolano, Edwin R. Hancock, Miguel A. Lozano, and Manuel Curado. The mutual information between graphs. *Pattern Recognition Letters*, 87:12–19, 2017.

[53] Soumik Purkayastha and Peter X.K. Song. Asymmetric predictability in causal discovery: an information theoretic approach. *arXiv preprint arXiv:2210.14455*, 2022.

[54] Ziya Zhang, Yi Li, Chen Xinguo, Yanzi Wang, Ben Niu, De Li Liu, Jianqiang He, Bakhtiyor Pulatov, Ishtiaq Hassan, and Qingtao Meng. Impact of climate change and planting date shifts on growth and yields of double cropping rice in southeastern China in future. *Agricultural Systems*, 205:103581, 2023.

[55] Robin A. A. Ince, Katarzyna Jaworska, Joachim Gross, Stefano Panzeri, Nicola J. van Rijsbergen, Guillaume A. Rousselet, and Philippe G. Schyns. The deceptively simple N170 reflects network information processing mechanisms involving visual feature coding and transfer across hemispheres. *Cerebral Cortex*, 26(11):4123–4135, 2016.

[56] Robin A.A. Ince, Bruno L. Giordano, Christoph Kayser, Guillaume A. Rousselet, Joachim Gross, and Philippe G. Schyns. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping*, 38(3):1541–1573, 2017.

[57] Stephanie J. Kayser, Robin A.A. Ince, Joachim Gross, and Christoph Kayser. Irregular speech rate dissociates auditory cortical entrainment, evoked responses, and frontal alpha. *The Journal of Neuroscience*, 35(44):14691–14701, 2015.

[58] Etienne Combrisson, Michele Allegra, Ruggero Basanisi, Robin A.A. Ince, Bruno Giordano, Julien Bastin, and Andrea Brovelli. Group-level infer-

ChinaXiv:202105.00070v16

ence of information-based measures for the analyses of cognitive brain networks from neurophysiological data. *NeuroImage*, page 119347, 2022.

[59] 汪方毅, 唐杰庆, 刘倩, 余成新, 李博, and 丁帆. 基于静息态 fMRI 区分健康老年人认知水平的 MVPA 方法研究. *磁共振成像*, 14(6):18–25, 2023.

[60] Pieter De Clercq, Jonas Vanthornhout, Maaike Vandermosten, and Tom Francart. Beyond linear neural envelope tracking: a mutual information approach. *Journal of Neural Engineering*, 20(2):026007, 2023.

[61] Paolo Victor Redondo, Raphaël Huser, and Hernando Ombao. Measuring information transfer between nodes in a brain network through spectral transfer entropy. *arXiv preprint arXiv:2303.06384*, 2023.

[62] 吴亚婷, 佘青山, 高云园, 谭同才, and 范影乐. 多尺度肌间耦合网络分析. *生物医学工程学杂志*, 38(4):742–752, 2021.

[63] Yating Wu, Qingshan She, Hongan Wang, Yuliang Ma, Mingxu Sun, and Tao Shen. R-Vine copula mutual information for intermuscular coupling analysis. In *Proceedings of the 11th International Conference on Computer Engineering and Networks*, pages 526–534, 2022.

[64] David Ó' Reilly and Ioannis Delis. A network information theoretic framework to characterise muscle synergies in space and time. *Journal of Neural Engineering*, 19(1):016031, feb 2022.

[65] Shaojun Zhu, Jinhui Zhao, Yating Wu, and Qingshan She. Intermuscular coupling network analysis of upper limbs based on R-vine copula transfer entropy. *Mathematical Biosciences and Engineering*, 19(9):9437–9456, 2022.

[66] 金国美, 佘青山, 马玉良, 张建海, and 孙明旭. 基于小波包-copula 互信息的肌间耦合特性. *传感技术学报*, 35(10):1348–1353, 2022.

[67] Johannes Leugering and Gordon Pipa. A unifying framework of synaptic and intrinsic plasticity in neural populations. *Neural Computation*, 30(4):945–986, 2018.

[68] Ari Pakman, Amin Nejatbakhsh, Dar Gilboa, Abdullah Makkeh, Luca Mazzucato, Michael Wibral, and Elad Schneidman. Estimating the unique information of continuous variables in recurrent networks. *Advances in Neural Information Processing Systems*, 2021.

ChinaXiv:202105.00070v16

[69] Liesa Ravijts. Revealing temporal interactions around the heartbeat-evoked potential modulated by emotional perception. Master's thesis, Ghent Univeristy, 2019.

[70] Agata Charzyńska and Anna Gambin. Improvement of the k-NN entropy estimator with applications in systems biology. *Entropy*, 18(1):13, 2015.

[71] Farzaneh Farhangmehr, Daniel M. Tartakovsky, Parastou Sadatmousavi, Mano R. Maurya, and Shankar Subramaniam. An information-theoretic algorithm to data-driven genetic pathway interaction network reconstruction of dynamic systems. In *2013 IEEE International Conference on Bioinformatics and Biomedicine*, pages 214–217, 2013.

[72] Aleksander Wieczorek and Volker Roth. Causal compression. *arXiv preprint arXiv:1611.00261*, 2016.

[73] Qiang Wu and Dongxi Li. CRIA: An interactive gene selection algorithm for cancers prediction based on copy number variations. *Frontiers in Plant Science*, 13:839044, 2022.

[74] 吴强. 基于 Copula 熵的交互式基因选择算法及其在癌症预测中的应用. 硕士学位论文, 太原理工大学, 2022.

[75] Radko Mesiar and Ayyub Sheikhi. Nonlinear random forest classification, a copula-based approach. *Applied Sciences*, 11(15):15, 2021.

[76] Jian Ma. Copula entropy based variable selection for survival analysis. *arXiv preprint arXiv:2209.01561*, 2022.

[77] Yu Luo, Guangcan Xu, Hongyu Li, Tianju Ma, Zi Ye, and Zhaohui Li. Research on establishing corneal edema after phacoemulsification prediction model based on variable selection with copula entropy. *Journal of Clinical Medicine*, 12(4):1290, 2023.

[78] S.M. Sunoj and N. Unnikrishnan Nair. Survival copula entropy and dependence in bivariate distributions. *REVSTAT-Statistical Journal*, 2023.

[79] Jian Ma. Predicting MMSE score from finger-tapping measurement. In *Proceedings of 2021 Chinese Intelligent Automation Conference*, pages 294–304, 2022. See also bioRxiv 817338 (2019).

[80] 李润泽, 姚尧, 冯珂珂, 杨硕, 李佳丽, 程铁峰, 尹绍雅, and 徐桂芝. 重复经颅磁刺激改善帕金森病运动症状的脑功能网络分析. *生物化学与生物物理进展*, 50(1):126–134, 2023.

ChinaXiv:202105.00070v16

[81] Jian Ma. Predicting TUG score from gait characteristics based on video analysis and machine learning. *bioRxiv*, page 963686, 2020.

[82] Jian Ma. Associations between finger tapping, gait and fall risk with application to fall risk assessment. *arXiv preprint arXiv:2006.16648*, 2020.

[83] 张婷婷, 王楠, 周天彤, 王苏弘, and 邹凌. 基于 Couple 熵的抑郁症相干性反馈指标提取. 电子测量技术, 45(9):160–167, 2022.

[84] 张婷婷. 基于脑电的抑郁症识别及虚拟现实康复训练研究. 硕士学位论文, 常州大学, 2022.

[85] Qingsong Shan and Qianning Liu. Binary trees for dependence structure. *IEEE Access*, 8:150989–150998, 2020.

[86] 罗良清, 平卫英, 单青松, and 王佳. 中国贫困治理经验总结：扶贫政策能够实现有效增收吗？. 管理世界, 38(2):70–83, 2022.

[87] Leonie Bossemeyer. Machine learning for causal discovery with applications in economics. Master's thesis, Ludwig-Maximilians-Universität München, 2021.

[88] 韦颖璐. 基于 pair–copula 熵的相关性度量. 硕士学位论文, 苏州大学, 2021.

[89] Muye Han and Jinsheng Zhou. Multi-scale characteristics of investor sentiment transmission based on wavelet, transfer entropy and network analysis. *Entropy*, 24(12):1786, 2022.

[90] Wuyue An, Lin Wang, and Dongfeng Zhang. Comprehensive commodity price forecasting framework using text mining methods. *Journal of Forecasting*, 2023.

[91] Yu-Xin Tian and Chuan Zhang. An end-to-end deep learning model for solving data-driven newsvendor problem with accessibility to textual word-of-mouth data. *SSRN*, 2023.

[92] 柳琼. 基于 Copula 和 MI 理论的相关性度量及其应用研究. 硕士学位论文, 三峡大学, 2018.

[93] 陈作海, 钱恒, and 高永超. 一种基于知识图谱的城市热线派单方法及系统, 2023. CN115860436A.

ChinaXiv:202105.00070v16

[94] Bowen Zhang, Jinping Lin, Man Luo, Changxian Zeng, Jiajia Feng, Meiqi Zhou, and Fuying Deng. Changes in public sentiment under the background of major emergencies – taking the Shanghai epidemic as an example. *International Journal of Environmental Research and Public Health*, 19(19):12594, 2022.

[95] Mario Wieser. *Learning Invariant Representations for Deep Latent Variable Models*. PhD thesis, University of Basel, 2020.

[96] Stuart William Card. Towards an information theoretic framework for evolutionary learning. Master's thesis, Syracuse University, 2011.

[97] 张可, 刘施彤, 郑植, 贾宇明, and 黄乐天. 一种基于动态贝叶斯网络的目标意图识别方法, 2022. CN114997306A.

[98] 许海云, 王超, 陈亮, 徐硕, 杨冠灿, and 朱礼军. 颠覆性技术的科学-技术-产业互动模式识别与分析. 情报学报, 42(7):816–831, 2023.

[99] Xueqian Fu, Hongbin Sun, Qinglai Guo, Zhaoguang Pan, Wen Xiong, and Li Wang. Uncertainty analysis of an integrated energy system based on information theory. *Energy*, 122(122):649–662, 2017.

[100] 朱正林 and 张冕. 基于 AO 优化 VMD-CE-BiGRU 的光伏发电功率预测. 国外电子测量技术, 41(10):56–61, 2022.

[101] 崔双双 and 孙单勋. 分工况下风电机组各变量相关性研究. 综合智慧能源, 44(12):49–55, 2022.

[102] 董海艳, 赵炳文, 王运韬, 田宇, 傅彦博, 孟德群, and 张铁. 一种含源荷时序相似度约束的源储协同规划配置方法, 2022. CN110766314A.

[103] Peili Liu, Song Han, Na Rong, and Junqiu Fan. Frequency Stability Prediction of Power Systems Using Vision Transformer and Copula Entropy. *Entropy*, 24(8):1165, 2022.

[104] 冯双, 杨浩, 雷家兴, 汤奕, 周吉, 钱俊良, and 郝珊珊. 一种电力系统宽频振荡影响因素和传播路径分析方法, 2022. CN114977222A.

[105] Wei Hu, Qiuting Guo, Wei Wang, Weiheng Wang, and Shuhong Song. Research on user loss contribution calculation of high-loss distribution area based on transfer entropy. In *2022 China International Conference on Electricity Distribution (CICED)*, pages 499–502, 2022.

[106] Xiaoping Xiong and Guohua Qing. A hybrid day-ahead electricity price forecasting framework based on time series. *Energy*, page 126099, 2022.

ChinaXiv:202105.00070v16

[107] Jiabei He and Lifeng Wu. Cross-conditions capacity estimation of lithium-ion battery with constrained adversarial domain adaptation. *Energy*, 277:127559, 2023.

[108] Marvin Lasserre, Régis Lebrun, and Pierre-Henri Wuillemin. Learning Continuous High-Dimensional Models using Mutual Information and Copula Bayesian Networks. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pages 12139–12146. AAAI Press, 2021.

[109] Marvin Lasserre. *Apprentissages dans les réseaux bayésiens à base de copules non-paramétriques*. PhD thesis, Sorbonne Université, 2022.

[110] Zhiwei Li, Peng Wang, Jili Zhang, and Hua Guan. A model-free method for identifying time-delay characteristics of HVAC system based on multivariate transfer entropy. *Building and Environment*, 217:109072, 2022.

[111] 曹久慧, 卞桂荣, 陈剑, 李宝枝, 顾晓彬, 徐士月, 夏显文, and 张旭辉. 一种基于自注意力机制的变形监测方法, 2023. CN116378120A.

[112] 黄达. 基于模块链构建的大件货物多式联运方案研究. 博士学位论文, 北京交通大学, 2021.

[113] 许罗豪, 刘金鑫, 张慧波, and 纪超. 基于熵与回归树的票价影响因素研究. 综合运输, 45(6):125–130, 2023.

[114] Yan-Ning Sun, Yu Chen, Wu-Yin Wang, Hong-Wei Xu, and Wei Qin. Modelling and prediction of injection molding process using copula entropy and multi-output SVR. In *IEEE 17th International Conference on Automation Science and Engineering*, 2021.

[115] 王小巧. 复杂机械产品装配过程质量自适应控制方法及支持系统研究. 博士学位论文, 合肥工业大学, 2015.

[116] Jie Dong, Keren Cao, and Kaixiang Peng. Hierarchical causal graph-based fault root cause diagnosis and propagation path identification for complex industrial process monitoring. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023.

[117] Jie Hu, Min Wu, Weihua Cao, and Witold Pedrycz. Dynamic modeling framework based on automatic identification of operating conditions for sintering carbon consumption prediction. *IEEE Transactions on Industrial Electronics*, pages 1–9, 2023.

ChinaXiv:202105.00070v16

ChinaXiv:202105.00070v16

[118] Fuqiang Sun, Wendi Zhang, Ning Wang, and Wei Zhang. A copula entropy approach to dependence measurement for multiple degradation processes. *Entropy*, 21(8):724, 2019.

[119] Min Yin, Jince Li, and Hongguang Li. A CNN approach based on correlation metrics to chemical process fault classifications with limited labeled data. *The Canadian Journal of Chemical Engineering*, 2022.

[120] Yingpeng Wei and Li Wang. Copula entropy-based PCA method and application in process monitoring. In *2022 4th International Conference on Intelligent Information Processing (IIP)*, pages 61–64, 2022.

[121] Xiaotian Bi, Deyang Wu, Daoxiong Xie, Huawei Ye, and Jinsong Zhao. Large-scale chemical process causal discovery from big data with Transformer-based deep learning. *Process Safety and Environmental Protection*, 2023.

[122] Baby Alpettiyil Krishnankutty, Rajesh Ganapathy, and Paduthol Godan Sankaran. Non-parametric estimation of copula based mutual information. *Communications in Statistics - Theory and Methods*, 49(6):1513–1527, 2020.

[123] Hao Liu, Dechang Pi, Shuyuan Qiu, Xixuan Wang, and Chang Guo. Data-driven identification model for associated fault propagation path. *Measurement*, 188:110628, 2022.

[124] Zefan Zeng, Guang Jin, Chi Xu, Siya Chen, Zhelong Zeng, and Lu Zhang. Satellite telemetry data anomaly detection using causal network and feature-attention-based LSTM. *IEEE Transactions on Instrumentation and Measurement*, 71:1–21, 2022.

[125] 吴格, 陈旭, 傅之凤, 李忠虎, and 杨程屹. 一种因果关系分析方法及装置, 2020. CN110766314A.

[126] Sheng Gao, Linchuan Zhang, Lei He, Xiaoyang Deng, Huilin Yin, and Hao Zhang. Attack detection for intelligent vehicles via CAN- bus: A lightweight image network approach. *IEEE Transactions on Vehicular Technology*, pages 1–13, 2023.

[127] 刘勃. 基于机器学习的封装材料加速预测. 硕士学位论文, 哈尔滨理工大学, 2022.

[128] Xu Wang, Liang Jin, Kaizhi Huang, Mingliang Li, and Yi Ming. Physical layer secret key capacity using correlated wireless channel samples. In

2016 IEEE Global Communications Conference (GLOBECOM), pages 1–6, 2016.

[129] 傅宇舟, 程文驰, 陈小军, and 李赞. 面向 6G 网络的基于语义通信的端到端服务框架. 移动通信, 47(6):35–40, 2023.

[130] Andreas Gocht-Zech. *Ein Framework zur Optimierung der Energieeffizienz von HPC-Anwendungen auf der Basis von Machine-Learning-Methoden*. PhD thesis, Technische Universität Dresden, 2022.

[131] Xuexing Zeng and T S Durrani. Band selection for hyperspectral images using copulas-based mutual information. In *2009 IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 341–344, 2009.

[132] Hudson and Thames. Machine learning financial laboratory (MLFinLab). GitHub, 2021. URL: https://github.com/hudson-and-thames/mlfinlab.

[133] Qiutong Wang. Social networks, asset allocation and portfolio diversification. Master's thesis, University of Waterloo, 2015.

[134] 朱仲儿. 多种机器学习方法的股票分类预测. 硕士学位论文, 上海师范大学, 2022.

[135] Rafael Calsaverini and Renato Vicente. An information-theoretic approach to statistical dependence: Copula information. *EPL (Europhysics Letters)*, 88(6):68003, 2009.

[136] Rafael S. Calsaverini. *Tópicos em Mecânica Estatística de Sistemas Complexos*. PhD thesis, Universidade de São Paulo, 2013.

[137] Fadhah Amer Alanazi. Truncating Regular Vine Copula Based on Mutual Information: An Efficient Parsimonious Model for High-Dimensional Data. *Mathematical Problems in Engineering*, 2021:4347957, October 2021.

[138] 熊靖宇. 基于 Copula 熵的行业风险溢出效应分析. 硕士学位论文, 东北财经大学, 2020.

[139] Mengyuan Chen, Jilan Liu, Ning Zhang, and Yichao Zheng. Vulnerability analysis method based on network and copula entropy. *Preprints*, 2023.

[140] 孔祥永, 王浩, 袁伟, and 蔡明. 一种自动化特征工程信用风险评价系统及方法, 2021. CN114049198A.

ChinaXiv:202105.00070v16

[141] Jian Ma. copent: Estimating copula entropy and transfer entropy in R. *arXiv preprint arXiv:2005.14025*, 2021.

[142] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69(6):66138, 2004.

[143] C.K. Chow and C.N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, 1968.

[144] Arthur Asuncion and David Newman. UCI machine learning repository, 2007.

[145] National Center for Health Statistics (NCHS). National Health and Nutrition Examination Survey, 2013-2014.

[146] Edward I. George. The variable selection problem. *Journal of the American Statistical Association*, 95(452):1304–1308, 2000.

[147] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[148] Hirotugu Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[149] Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

[150] Robert Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society Series B-Methodological*, 58(1):267–288, 1996.

[151] Arthur E. Hoerl and Robert W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–86, 2000.

[152] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 67(2):301–320, 2005.

[153] Arthur Gretton, Kenji Fukumizu, Choon H. Teo, Le Song, Bernhard Schölkopf, and Alex J. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20*, volume 20, pages 585–592, 2007.

ChinaXiv:202105.00070v16

[154] Niklas Pfister, Peter Bühlmann, Bernhard Schölkopf, and Jonas Peters. Kernel-based tests for joint independence. *Journal of The Royal Statistical Society Series B-statistical Methodology*, 80(1):5–31, 2018.

[155] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.

[156] Gábor J. Székely and Maria L. Rizzo. Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265, 2009.

[157] Hui Zou. The adaptive LASSO and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

[158] Ruth Heller, Yair Heller, Shachar Kaufman, Barak Brill, and Malka Gorfine. Consistent distribution-free K-sample and independence tests for univariate random variables. *Journal of Machine Learning Research*, 17(1):978–1031, 2016.

[159] Wassily Hoeffding. A non-parametric test of independence. *Annals of Mathematical Statistics*, 19(4):546–557, 1948.

[160] Wicher Bergsma and Angelos Dassios. A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, 20(2):1006–1028, 2014.

[161] Wenliang Pan, Xueqin Wang, Heping Zhang, Hongtu Zhu, and Jin Zhu. Ball covariance: A generic measure of dependence in banach space. *Journal of the American Statistical Association*, 115(529):307–317, 2020.

[162] Norbert Wiener. The theory of prediction. modern mathematics for engineers. *New York*, 165, 1956.

[163] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

[164] Clive WJ Granger. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control*, 2:329–352, 1980.

[165] Thomas Schreiber. Measuring information transfer. *Physical Review Letters*, 85(2):461–464, 2000.

ChinaXiv:202105.00070v16

[166] Xuan Liang, Tao Zou, Bin Guo, Shuo Li, Haozhe Zhang, Shuyi Zhang, Hui Huang, and Song Xi Chen. Assessing Beijing's PM2.5 pollution: severity, weather impact, APEC and winter heating. *Proceedings of The Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257, 2015.

[167] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *UAI'11 Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 804–813, 2011.

[168] Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. Conditional distance correlation. *Journal of the American Statistical Association*, 110(512):1726–1734, 2015.

[169] Alfréd Rényi. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica*, 10:441–451, 1959.

[170] B. Schweizer and E. F. Wolff. On Nonparametric Measures of Dependence for Random Variables. *The Annals of Statistics*, 9(4):879 – 885, 1981.

[171] Jian Ma. Evaluating independence and conditional independence measures. *arXiv preprint arXiv:2205.07253*, 2022.

[172] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938.

[173] Wolfgang Trutschnig. On a strong metric on the space of copulas and its induced dependence measure. *Journal of Mathematical Analysis and Applications*, 384(2):690–705, 2011.

[174] Kai Zhang. BET on independence. *Journal of the American Statistical Association*, 114(528):1620–1637, 2019.

[175] C Genest, J G Nešlehová, B Rémillard, and O A Murphy. Testing for independence in arbitrary distributions. *Biometrika*, 106(1):47–68, 01 2019.

[176] Arturo Erdely. A subcopula based dependence measure. *Kybernetika*, 53(2):231–243, 2017.

[177] Xiaofeng Shao and Jingsi Zhang. Martingale difference correlation and its use in high-dimensional variable screening. *Journal of the American Statistical Association*, 109(507):1302–1318, 2014.

ChinaXiv:202105.00070v16

[178] Trevor Park, Xiaofeng Shao, and Shun Yao. Partial martingale difference correlation. *Electronic Journal of Statistics*, 9(1):1492 – 1517, 2015.

[179] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021.

[180] Mona Azadkia and Sourav Chatterjee. A simple measure of conditional dependence. *The Annals of Statistics*, 49(6):3070 – 3102, 2021.

[181] Fred Viole and David N. Nawrocki. Deriving nonlinear correlation coefficients from partial moments. *SSRN*, 2012.

[182] Eric V. Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1):20180017, 2019.

[183] Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514 – 1538, 2020.

[184] Cyrill Scheidegger, Julia Hörrmann, and Peter Bühlmann. The weighted generalised covariance measure. *arXiv preprint arXiv:2111.04361*, 2021.

[185] Zhen Huang, Nabarun Deb, and Bodhisattva Sen. Kernel partial correlation coefficient – a measure of conditional dependence. *arXiv preprint arXiv:2012.14804*, 2020.

[186] Lasse Petersen and Niels Richard Hansen. Testing conditional independence via quantile regression based partial copulas. *Journal of Machine Learning Research*, 22(70):1–47, 2021.

[187] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947, 2017.

[188] Octavio César Mesner and Cosma Rohilla Shalizi. Conditional mutual information estimation for mixed, discrete and continuous data. *IEEE Transaction on Information Theory*, 67(1):464–484, jan 2021.

[189] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Fast conditional independence test for vector variables with large sample sizes. *arXiv preprint arXiv:1804.02747*, 2018.

ChinaXiv:202105.00070v16

[190] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[191] Samuel Burkart and Franz J Király. Predictive independence testing, predictive conditional independence testing, and predictive graphical modelling. *arXiv preprint arXiv:1711.05869*, 2017.

[192] Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022–140022, 2014.

[193] Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Research International*, 2014:781670, 2014.

[194] Florian Hodel. cylcop: Circular-linear copulas with angular symmetry for movement data. CRAN, 2022. R package version 0.2.0, URL: `https://cran.r-project.org/package=cylcop`.

[195] Kai Xu. The CopEnt.jl package. Github, 2021. URL: `https://github.com/xukai92/CopEnt.jl`.

[196] JuliaDynamics. The CausalityTools.jl package. Github, 2023. URL: `https://github.com/JuliaDynamics/CausalityTools.jl`.

[197] Robin A.A. Ince, Jan-Mathijs Schoffelen, Lukas Snoek, and Danylo Ulianych. gcmi : Gaussian-Copula Mutual Information. GitHub, 2020. URL: `https://github.com/robince/gcmi`.

ChinaXiv:202105.00070v16